



The Graduate Institute of Science and Engineering

M.Sc. Thesis in Electrical and Computer Engineering

EVALUATION OF SEMANTIC WEB SEARCH ENGINES

by

Farouk Musa ALIYU

June 2014
Kayseri, Turkey

EVALUATION OF SEMANTIC WEB SEARCH ENGINES

by

Farouk Musa ALIYU

A Thesis submitted to
The Graduate institute of Science and Engineering

of

Melikşah University

in partial fulfillment of the requirement for the degree of
Master of Science

in

Electrical and Computer Engineering

June 2014
Kayseri, Turkey.

This is to certify that I have read the thesis entitled “Evaluation of Semantic Search Engines” by Farouk Musa Aliyu and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Electrical and Computer Engineering, the Graduate Institute of Science and Engineering, Melikşah University.

June 25, 2014

Assoc. Prof. Dr. Ahmet Uyar
Supervisor

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

June 25, 2014

Prof. Dr. Murat Uzam
Head of Department

Examining Committee Members

Assoc. Prof. Dr. Ahmet Uyar

June 25, 2014

Asst. Prof. Dr. Kadir A. Peker

June 25, 2014

Asst. Prof. Dr. Mehmet S. Aktaş

June 25, 2014

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Science and Engineering.

Prof. Dr. M. H. Keleştemur
Director

June 2014

EVALUATION OF SEMANTIC WEB SEARCH ENGINES

Farouk Musa ALIYU

MSc Thesis- Electrical and Computer Engineering
June 2014

Supervisor: Assoc. Prof. Dr. Ahmet UYAR

ABSTRACT

On their effort to improve their web search engines by providing direct answers to many user queries and eradicates the problems in dealing with ambiguous queries, Google and Bing developed semantic web search engines. They have built a huge database of entities and relationships among them. Their objective is to have a comprehensive entity database. They want to have all entities that may be of an interest to the users. However, this is a challenging task, since the entity databases could not be built without human intervention. Both search engines are constantly improving their semantic search engines and adding new features. The evolution of these semantic search engines has really changed the normal routine operation of web search engines of returning results based on lexical similarity or keyword matching. This is a step ahead of the state-of-the-art web search engines.

In this study, we investigated the coverage of classes in Google Knowledge Graph and Bing's Satori, how they organized their data, the extent to which they answer user's queries about list of entities, kinds of queries supported by the engines and the extent to which they understand natural language queries.

Keywords: Search engine evaluation, semantic search engines, entity databases

ANLAMSAL WEB ARAMA MOTORLARININ DEĞERLENDİRİLMESİ

Farouk Musa ALIYU

Yüksek Lisans Tezi- Elektrik ve Bilgisayar Mühendisliği Bölümü
June 2014

Tez Danışmanı: Doç. Dr. Ahmet UYAR

ÖZ

Kullanıcı sorgularına direk cevaplar verebilmek ve kullanıcı sorgularını daha iyi anlayıp daha doğru sonuçlar üretebilmek için, Google ve Bing arama motorları anlamsal arama motorlarını geliştirmişlerdir. Bunun için büyük bir varlık-ilişki veri tabanı geliştirmişlerdir. Hedefleri kullanıcılar için ilgili olabilecek her konuda bu veri tabanında bilgi bulundurmaktadır. Fakat, bu veri tabanları insan müdahalesi olmadan otomatik olarak geliştirilememektedir. Her iki arama motoru da anlamsal arama motorlarını sürekli geliştirmekte ve yeni özellikler eklemektedir. Anlamsal arama motorları, sözlüksel benzerlik ve anahtar kelime karşılaştırması yapan web arama motorlarının normal çalışma sistemlerini çok önemli oranda etkilemiştir. Anlamsal arama motorları gelecekteki daha kapsamlı arama motorları için önemli bir adım oluşturmaktadır.

Bu tezde, Google Knowledge Graph ve Bing Satori anlamsal arama motorlarının sınıf kapsamını, varlık ilişki yöntemlerini, kullanıcı sorgularına varlık listesi döndürme özelliklerini, kullanıcıların İngilizce sorgularını anlama kapasitelerini araştırdık.

Anahtar Kelimeler: Arama motordur değerlendirme, anlamsal arama motorları, varlık veri tabanlar

DEDICATION

This thesis is dedicated to the Kano State Governor (Eng. Dr. Rabi'u Musa Kwankwaso) and to the family of Alhaji Musa Aliyu Rano..

AKCNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Ahmet UYAR Without whom this thesis wouldn't have been completed. I am also grateful for his upmost support contributions, care and advice.

Particular appreciation to the Kano state Governor Eng. Dr. Rabi'u Musa Kwankwoso for the opportunity he has given us to study in a prestigious University.

I would like to acknowledge and thank all my friends, family and my fellow master's colleagues for their advice and guidance throughout our studies.

TABLE OF CONTENTS

ABSTRACT.....	iii
ÖZ	iv
DEDICATION.....	v
AKCNOWLEDGEMENT.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS.....	xiii
CHAPTER 1 INTRODUCTION.....	1
1.1 Motivation.....	2
1.2 Advantages Of Semantic Web Search Engines	3
1.3 Research Questions	4
1.5 Thesis Organization	6
CHAPTER 2 LITERATURE REVIEW	7
2.1 Web Search	7
2.2 Semantic Search.....	9
2.3 Semantic Search Engines	10
2.4 Search Engine Evaluation Studies	12
2.5 Related Work	13
CHAPTER 3 GKG AND SATORI ENTITY TYPES AND THEIR HIERARCHY	14
3.1 Introduction.....	14
3.2 Methodology	16
3.2.1 Data Collection (Classes and Entities)	17
3.2.2 Search Engine Settings	20
3.2.3 Testing the Entities	20

3.3 Experimental Results	21
3.4 Hierarchical Structure Of Entity Types.....	26
3.4.1 Hierarchy in GKG	27
3.4.2 Hierarchy in Satori	29
3.5 Features of Entities Displayed In Gkg And Satori.....	30
3.6 Conclusion.....	31
CHAPTER 4 LIST SEARCH SERVICES FOR ENTITIES	34
4.1 Introduction	34
4.1.1 Research Question	35
4.1.2 Multiple Entity Search Result	36
4.1.2.1 Simple carousel result.....	39
4.1.2.2 Complex carousel	41
4.2 Methodology	41
4.2.1 Class Selection.....	42
4.2.3 Search Engine Settings	45
4.2.4 Query Testing	45
4.2.5 Relevance Judgment.....	45
4.3 Results	46
4.3.1 GKG Result	46
4.4 Conclusions	49
CHAPTER 5 QUERIES SUPPORTED BY GKG AND SATORI.....	50
5.1 Introduction	50
5.1.1 Research Questions	50
5.2 Query Interfaces Of Semantic Search Engines	51
5.2.1 Structured query languages	51
5.2.2 Form-based	52
5.2.3 View based	52
5.2.4 Natural Language Queries	53
5.3.1 Single Entity	54
5.3.2 Set of Entities	55
5.3.3 Attribute of Entity.....	55

5.4 Objective of The Chapter	56
5.5 Methodology	57
5.5.1 Search Engine Settings	58
5.5.2 Query Sets	58
5.5.2.1 Geo-Queries	59
5.5.2.2 Real Query log	60
5.5.3 Query Types	60
5.5.4 Evaluation Scale	62
5.5.5 Categorization of Geo-Queries Based On Their Complexities	64
5.5.6 Categorization of Yahoo Queries	65
5.6 Experimental Results:	67
5.6.1 Geo-queries Result	67
5.6.1.1 GKG Result.....	68
5.6.1.2 Satori Result.....	70
5.6.2 Geo-Queries Result Based On Their Complexities.....	72
5.6.2.1 GKG Result.....	72
5.6.2.2 Satori Result.....	77
5.6.3 Yahoo Queries Result.....	82
5.6.3.1 GKG Result.....	82
5.6.3.2 Satori Result.....	83
5.6.4 Yahoo Query Result For List Search.....	84
5.7 Conclusions	85
CHAPTER 6 CONCLUSIONS AND FUTURE WORK.....	88
6.1 Conclutions	88
6.2. Future Work	91
REFERENCES	93

LIST OF TABLES

Table 4.1: Query types and some sample queries.....	44
Table 5.1: The number of queries in each query type for the Geo-queries queries.....	62
Table 5.2: Interpretation of scales.....	64
Table 5.3: Number of queries filtered from each query categories.	67

LIST OF FIGURES

Figure 3.1: percentage of classes presents in GKG and Satori from our class set.....	22
Figure 3.2: GKG result showing higher and lower classification.....	28
Figure 3.3: GKG result not showing lower and higher classification for searching cat.....	29
Figure 3.4: searching for panthera in Satori	30
Figure 4.1: Example of list search results in GKG.....	37
Figure 4.2: searching for ‘tourist attractions in turkey’ in GKG.....	38
Figure 4.3: Example of list search results in Satori	38
Figure 4.4: Carousel with double filters.....	39
Figure 4.5: carousel result with single filter in GKG.....	40
Figure 4.6: complex carousel in GKG.....	41
Figure 4.7:GKG result for multiple entity search.....	47
Figure 4.8: Satori result for multiple entity search.....	48
Figure 5.1: Single entity result.....	54
Figure 5.2: multiple entity result in Satori.....	55
Figure 5.3: Attribute of an entity result in GKG.....	56
Figure 5.4: Geo-queries and the results in GKG and Satori.....	68
Figure 5.5: GKG Correct and incorrect result for each query type from Geo-queries.....	69
Figure 5.6: Satori Correct and incorrect result for each query type from Geo-queries.....	71
Figure 5.7: GKG result for single entity type from the Geo-queries.....	73
Figure 5.8: GKG result for multiple entities type from the Geo-queries.....	74
Figure 5.9: GKG result for attribute of an entity type from the Geo-queries.....	75
Figure 5.10: GKG result for attribute of multiple entities type from the Geo-queries.....	76
Figure 5.11: GKG result for statistical queries from the Geo-queries.....	76
Figure 5.12: GKG result with respect to query complexities.....	77
Figure 5.13: Satori result for single entity type from the Geo-queries.....	78
Figure 5.14: Satori result for multiple entities type from the Geo-queries.....	79

Figure 5.15: Satori result for attribute of an entity type from the Geo-queries.	79
Figure 5.16: Satori result for attribute of multiple entities type from the Geo-queries.	80
Figure 5.17: Satori result for statistical queries from the Geo-queries.	81
Figure 5.18: Satori result based on query complexities for combine query types.	82
Figure 5.19: GKG result for the Yahoo queries.....	83
Figure 5.20: Satori result for the Yahoo queries.....	84

LIST OF ABBREVIATIONS

GKG: Google Knowledge Graph

SERP: Search Engines Result Page

URL: Uniform Resource Locator

RDF: Resource Description Framework

OWL: Web Ontology Language

SPARQL: SPARQL Protocol and RDF Query Language

CHAPTER 1

INTRODUCTION

Today, major search engines are in the process of an important technological shift. Traditionally they have been linking users to documents on the web by returning a list of documents to the user queries. They were acting as an intermediary between users and the documents on the public web. Although this will continue to be an important component of search engines, lately they are moving to become fully semantic search engines. They want to understand the user queries semantically and serve their information needs precisely from their knowledge repositories. They want to answer many of the user information needs directly. They are working on to build large knowledge repositories about real world entities and concepts. Google calls its entity database as knowledge graph and introduced it in May of 2012 [15]. Bing calls its entity database as Satori and introduced it in June of 2012 [16].

Both Google and Bing are developing entity databases for hundreds of millions of entities. These entities are not documents on the web, but rather constructed information about real world objects and concepts including people, places, books, movies, events, arts, science, etc. An entity may have some properties and relationships to other entities. The relationships of entities are particularly important. They turn the entity database into a graph. Initially, Google reported to have more than 500 million entities and more than 3.5 billion relationships [15]. Similarly, Bing is planning to build a database with billions of entities and relationships [16]. However, building an entity database is a challenging task and continuing process.

Entity databases are usually implemented as a graph database to better handle the connections among them [27]. Therefore, they are very different from the traditionally used

inverted index file systems in search engines. This requires search engines to redesign many of the previously used algorithms for entity databases. New relevancy detection and entity ranking mechanisms are needed. Traditionally used ranking algorithms such as PageRank [24], [49], HITS [25], SpamRank [26], etc. need to be reformulated.

Building an accurate and comprehensive entity databases is a challenging task [19]. It is much harder to build them compared to the traditional webpage corpuses. When building a webpage corpus, crawlers discover the existing web pages. They surf the web by following http links and download the ones that may be of an interest to the users [24]. However, building entity databases involves creating entities and relationships among them. It is not a task of discovering existing documents but rather building entities from the scratch. Search engines use both automatic data extraction algorithms and human workers to build entity databases [17], [18]. They use public resources on the web such as Wikipedia and social media, government organization datasets such as CIA World Factbook, digital book contents such as Google Books, etc.

Since it is a time consuming and difficult task to build entity databases, it is important for search engines to cover entities that are most helpful to its users. Both Google and Bing report that they are primary motivated by the user search query logs and build entities for objects that are searched most. Google started with: “landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art and more” [15]. Initially Google did not have “hotels, restaurants, corporations, and events” as entities [18]. However, they now cover these kinds of objects. Bing Satori started with three types of entities: “movies, restaurants and hotels” in June 2012. Soon, they expanded the list with “people, places, and things”. They report that these are the most commonly searched entities on Bing and about 10 percent of all user queries are for “people” searches [16].

1.1 MOTIVATION

Most traditional search engines have been serving result to user’s queries based on matching queries or keywords to a large database of documents. They first crawl the web and index the documents they find useful, they process user’s query and return those

documents that are most relevant to the query. Thousands or millions of documents usually matched to the user queries, most of which may not be relevant to the users' intent. Therefore, they try to list documents that may be most relevant. Some years back, indexing a lot of documents has been one of the major challenges of search engines. Conversely, in the current era the major challenge has been providing the most relevant results in a simple and friendly interface. This has indeed created a major competition among search engines. Traditional information search engines have a major weakness, which is their inability to handle ambiguous queries. However, with the increasing availability of structured data on the web, semantic search is increasingly becoming popular to eradicate this problem.

1.2 ADVANTAGES OF SEMANTIC WEB SEARCH ENGINES

Determination of User Intention: traditional search engines have difficulty resolving ambiguous queries (queries that have multiple meaning). This is because they match keywords to documents; they do not understand the meaning of what the user is searching for. For queries including ambiguous words like 'windows', they could not understand whether the user is searching for Microsoft windows or windows for buildings. If a mixed of result is provided, this will of course reduces precision and recall with respect to what is taken to be a relevant result or the user intent. Therefore, there is a need to disambiguate this type of queries in order to maximize system performance. Semantic search uses the context and meaning of the sentence of phrases pass by user's to understand their intent. This will help in providing users with the most relevant answers to their queries hence increasing precision and recall.

Discovery of Related Information: One advantage of semantic search engines is the discovery of related information by users about their search needs. This allows users to go deeper and wider to accumulate more knowledge about what they are interested in. By modelling entities and their relationship, semantic search engines can understand how entities are connected or linked to other entities. This makes semantic search engines to better understand the real-world and suggest related entities users are likely to be interested. For example, for a search about the movie "Frozen", search engines can suggest other movies like The Lion King, Despicable Me, Shrek, or Ice Age. Probably because they are

the highest grossing computer animated films or they may suggest other entities people may be interested to know such as the movie casters.

Providing Direct Answers: Most web search engines return millions of documents or links for a single query. Moreover, users need to examine a lot of documents before they can actually find their search needs. They don't provide direct answer to user's queries about factual things such as: "who is the president of Turkey", "how big is Alaska", "how high is Mountain Everest", "who is the author of Dreams from my Father", "what cities are in California" etc. However, some semantic search engines like Google Knowledge graph and Bing's Satori seems to give directly answer to user's queries about factual things. This is very important since the search engines saves a lot user's time trying to get such kind of information or answers on the web.

1.3 RESEARCH QUESTIONS

The questions we hope to answer in our research include the following:

1. What kinds of entities do GKG and Satori cover and how they organize the entities?
2. How often do they answer user's queries about multiple entities?
3. What kinds of queries invoke semantic search component of the search engines?

We attend to each of the questions in chapter 3, 4 and 5 respectively. We analyze each question given the methods we used the results we obtained and the conclusion we are able to come up with.

In the first question, we would like to examine the kinds of classes indexed by the semantic search engines such as people, hotels, rivers, movies, theatres etc. Also we would like to examine the organization of their data, whether they implement hierarchies among their types and the level of the hierarchy. We aim to have a deeper understanding of the kinds of entities these semantic search engines are interested, how much and how they tend to organize and structure their data. This is a very important aspect of the semantic search engines and it will help to have a better understanding of the architecture of the semantic search engines and what they can accomplish with their data.

In the second question, we would like to investigate the list services offered by the semantic engines. That is, how much of the user's queries about list of entities can they answer? List search services by semantic search engines are very important. It allows the user to get information about related things or group of things so that they can learn and find interesting things about them. It also shows how well the search engines understand the real-world entities and their relationships. For example, searching for "cities in California" will allow users to learn more about California and come to know other cities in California that they might not know. Similarly, searching for queries like "2014 movies", "Chinese restaurants in Texas", "songs by Michael Jackson", "books by J. K. Rowling", "famous politicians in USA", "tourist attractions in Paris" etc. will allow users find new things and learn interesting things about group of entities.

In the third question, we would like to investigate the types of queries the semantic search engines can answer, their natural language understanding capability and how much they understand about the most commonly used user queries? These are important aspect to investigate. The kind of queries supported by the semantic engines will determine the capabilities or usefulness of the semantic engines to their users. Their natural language understanding will answer if the semantic engines can understand user's queries or sentences of various complexities.

1.4 CONTRIBUTIONS

The contribution of our research centers on the significance of search engine evaluations.

Our research can be used by search engine users to improve their search experience, it will enlighten them on how to make an efficient use of the search engines and allow them to know what are/are not supported by these search engines. In addition, researchers will also find our study very helpful especially on the result obtained from our evaluation. Our research will also motivate other researchers who are interested in semantic search.

1.5 THESIS ORGANIZATION

This thesis is divided into six (6) chapters. In the first chapter, we explained the overview and background of the proposed research. This includes: how major search engines are changing from information engines to knowledge engines, we introduce Google Knowledge graph and Being's Satori and how they tend to solve the problem of ambiguity that is facing the current traditional search engines. We state the objectives of our study and the contributions we made.

In the second chapter, we discuss the basic concepts of web search engines and how they work, semantic search, semantic search engines, search engines evaluation studies and related work.

Chapter three investigated the classes covered by the semantic search engines and the organization of entity types in their databases. We gave sample classes found in the search engines and those not indexed by the search engines. We stated the method we used in our investigation, the results we obtained and the conclusions we derived.

Chapter four focuses on list search services provided by the semantic search engines and type of multiple entity results. We state the methods we used, the results we obtained and the conclusions we derived.

In the fifth chapter, we discuss the kind of queries invoked by the search engines to give result from their entity database, we tested queries of different complexities including real queries and constructed queries, we state the methods we adopt, results and the conclusion we derived.

In chapter six we conclude the study by summarizing our findings and discussion of the possible future work.

CHAPTER 2

LITERATURE REVIEW

The objective of this chapter is to analyze an introductory frame work needed to understand the basis of our studies. In this chapter, we discuss web search, semantic search, semantic search engines and search engine evaluation studies. These are the background or foundation of our studies.

2.1 WEB SEARCH

Web search is simply searching for documents or webpages on the web (WWW). These documents or information resources may be of different forms or formats which may include text documents (word documents, text files, PDF etc.), html files, movies, sounds/music, pictures/images, emails etc. The documents on the web are typically unstructured. These types of documents can be understood by humans but machines need some additional mechanisms or interpretation to understand the documents. Software systems that are capable of crawling and indexing web documents to answer user's demands or queries are known as web search engines.

The history of web search engines dated back in the early 90s and has recorded some tremendous achievements until then. The early Search engines suffer from many challenges some of which includes storage space and index limitations, accuracy, efficiency, relevancy etc. Even though most of these problems are not a threat to major search engines of the present, providing relevant result has been a key consideration until today and it will continue to be a major influence in the scaling of search engines.

There are basically three processes involved in web search engines:

- (a) **Crawling:** Crawling is a process that spans or navigates the web using a robot or otherwise known as crawlers [46]. Crawlers download all the pages or documents they encounter across their path while crawling. However, there are several issues the search engines engaged to make crawlers more efficient. For this reason, the crawlers consist of complex algorithms [47] to avoid duplication and downloading documents that are up-to-date in the system. These make crawlers more robust. Crawlers may sometimes crash trying to download documents. Not all pages are crawl-able such as emails and some social sites. Crawlers may also go in a state of deadlock when they visited sites with no out-links. Search engines may employ several crawlers in different location and some may also be used as backups [48].
- (b) **Indexing:** indexing involves parsing documents and constructing an inverted index files for each distinct word. Each inverted index file contains the document ids of all documents having a particular word. For example, the inverted index file for the word “java” contains the ids of all documents that contain this word in the crawled file collection. In addition to the ids of documents, this file also contains other information such as the word position in the document, font size etc. Document ids are stored as sorted for speedy processing when performing the searches.

Obviously, indexing all the documents on the web into inverted files is not enough to output a qualitative result by matching keyword [48]. For the documents retrieved after matching keywords, search engines must employ some mechanism to return the top relevant results. Google use the advantage of the link structure of the web and anchors to return results with high precision [48]. They developed an algorithm called PageRank [49]. PageRank is an indirect measure of the relative importance of web page to people and it is an excellent way to distinguish web keyword results with respect to their importance or quality.

- (c) **Query processing:** One of the most delicate processes in search engine operation is query processing. The aim of processing user’s queries is to provide the most relevant result for his information needs. This has not been an easy task. It does not just include

providing the most relevant result but search engines need to operate efficiently and provide result in a more user friendly manner. Query processing starts by parsing user queries and retrieving the inverted index files for each word in the query. The documents which match the searched terms are retrieved by examining these inverted index files. The documents are then ranked according to their relevancy to the searched query. The top documents are presented to the user. Millions of documents may match a user query. However, only 10 documents are usually returned to the user in the first search engine result page (SERPs). Additional results may be retrieved by clicking through the next pages in SERP. Each result may contain a heading that describe the result, a URL that locate the document or page, and a brief summary of the document content to allow users to read from a part of the document before they click. Most users don't click on the results beyond the first page. This again shows why it is important to return the most relevant results in the first page. Several processes are involved in ranking the documents retrieved by the search engines [48]. Each search engine may employ its own mechanisms. But the net result is to return the most relevant result to the user. Some of the mechanism involved in ranking documents retrieved by the web search engines may include the use of PageRank [49], [27], similarity scores [50], [51], etc. These figures are usually computed by considering the titles, anchors, URLs, and fonts in the documents.

2.2 SEMANTIC SEARCH

The goal of semantic search is to provide users more accurate answers to their queries by understanding the meaning and context of their queries. Most traditional search engines relied on matching keywords to a set of documents and using a ranking algorithm such as PageRank to provide their users with relevant results. Most of their search processes are controlled by machines which are usually made of robust programs. However, these machines operate “blindly” because they don't know the meaning of what the users are searching [53]. For example searching for ‘Michael Jackson’ in tradition search engines may provide relevant results, not because they know Michael Jackson as a person and famous singer but because the query words is mentioned in the document (the keyword ‘Michael Jackson’ is fortunately the same as the entity name). They only need to be reliable

about the source and score the documents to return the top ranked. Sometimes, these search engines are even fooled by site owners who have less information about what the user is searching but want their page to be returned at the top of the SERP. A major problem with these kinds of search engines is that they find it difficult to handle user's queries that are ambiguous [53]. Because they are not sure of what the user is actually seeking for, they usually return millions of documents per query whereas users only use a little out of the result.

Semantic search seems to be the solution of most traditional search engine's problems. They are bound to provide more relevant and reliable results than traditional search engines if properly implemented. [34] Evaluate three web search engines (Google, Yahoo and Msn) by comparing their result to a semantic search engine (Hakia). Another study [53] also shows that adding semantic to web search engines produces result better than the top web search engines such as Google. Semantic search engines use the data available in semantic web to provide the most relevant results to their users [43]. They answer user's needs by interpreting their queries against knowledge. By gathering a large amount of knowledge database, search engines can use pattern matching [63], [77] and query interpretation [78] to identify the entities and concepts in user's queries and replace them with actual or more specific entities to refine the queries to a more promising one [53]. The result obtained from the refine queries can then be ranked and return to the user. The aim is to provide users with the answers of their queries rather than outputting a bunch of loosely related results. Understanding user's need is the best solution to answer what they need. This is in accordance to the saying '*understanding a question is more than answering the question*'.

2.3 SEMANTIC SEARCH ENGINES

One of the major problems that hindered the scaling of semantic search engines is the low availability of structured data. Most website owners are yet to implement the semantic standards such as Schema.org [52] or other standards so that semantic search engines or ontology base system can be able to index their structured data. This is a major problem

that slows the growth of the semantic web even though the numbers of sites that are implementing the standards are increasing over the years.

The standards by Schema.org provide web masters with a collection of shared vocabularies for which they can use to mark up their web pages so that major search engines such as Google [1], Bing [3], Yahoo [2], Yandex [4] etc. can understand. Schema.org vocabularies and microdata format HTML content inside the HTML tags. This allows web masters to add more semantics to their page by identifying and defining the entities in their page. The scope of the entity is defined inside an opening tag using *itemscope*, and the closing tag will make the end of the entity scope. The *itemtype* (entity type) is defined after the *itemscope*. The value of *itemtype* is usually a URL such as "http://schema.org/Movie". Next the properties of the item are defined using *itemprop*. A sample code for the movie "Toy Story" is given below:

```
<div itemscope itemtype ="http://schema.org/Movie">
  <h1 itemprop="name"> Toy Story </h1>
  <span>Director: <span itemprop="director"> John Lasseter </span> (born
January 12, 1957)</span>
  <span itemprop="genre"> Animation </span>
  <a href="../../movies/Toy-Story-trailer.html"
itemprop="trailer">Trailer</a>
</div>
```

The *itemscope* is defined inside the <div> tag. Therefore, the scope of the item lies between opening: tag <div> and the closing tag: </div>. The item type is defined to be movie using the URL. Three properties of the movie Toy Story are defined which are director, genre and trailer. With this code, the search engines would be able to understand what Toy Story is, not just as a string of characters. Other properties about entities can be obtained from other sites and merge with the known properties.

Semantic search engines operations differ from traditional or web search engines. They do not index documents as keywords rather they gather a large amount of ontologies in RDF, OWL etc. or crowd the web to index entities and their relationship. With entities and their relationships, semantic search engines can understand the real-world. Entities are defined by properties or attributes and an entity belongs to a certain type or class. A class

denotes a concept. It consists of entities defined with similar properties. Semantic search engines are systems that take user's queries as an input, process the queries with respect to a set of ontologies or knowledge base and return the appropriate answer to the query. The inputs may be in the form of keywords, formal language such as SPARQL, forms, RDF triple, graph, natural language queries etc. The result of semantic search engines may take different forms. Result may be presented in the form of RDF triple describing the entities or as entities and their attributes.

Different semantic search engines have emerged, however the popular ones includes Hakia [5], Swoogle [6], GoPubMed [7], Kosmix [8] and the recent Google Knowledge Graph and Bing Satori etc.

2.4 SEARCH ENGINE EVALUATION STUDIES

Search engine evaluation has been the major driving force in the evolution and development of search engines [32]. Search engine evaluation is usually carryout to assess the features of semantic search engines which may include performance, output, users and uses, interfaces, algorithms etc. The significance of search engine evaluation is twofold: to help Web users in their choice of search engines and to inform the development of search algorithms and search engines [33]. Over the years, there have been a lot of search engine evaluations. Some proposed and test new evaluation methodology or measures for evaluating search engines such as [33], [39], [40], [43], others may evaluate one or more of the search engine's features including accuracy of result/output, simplicity of user's interface, accuracy of hit count, search algorithms etc. such as [42], [44], [45] . While others may not evaluate any feature system but may survey search engines from the top level to state their problems, possible solutions and some research directions such as [41], [37], [38]. However, must of these evaluations focuses on traditional or web search engines as a result of their popularity, large number of users and the bulk of data available for unstructured data on the web. Little research has been recorded on semantic search and semantic search engines compare to research on other IR systems, although the research trend is increasing recently [34].

2.5 RELATED WORK

To the best of our knowledge, this is the first evaluation studies that directly evaluate Google knowledge Graph and Bing's Satori. Since the enouncement of Google Knowledge Graph (GKG) and Bing's Satori in 2012, the only work we could find in the literature about either of these semantic search engines are [36] and [35], which are not related to our study. The first suggested a browser extension prototype which enhances the result outputted in GKG result panel by using the comments users made about real-world entities on social networks (Google+, Facebook and Twitter). While the later also use a browser extension to extract anonymous Google Knowledge Graph facts from their SERPs to populate an external data source aimed at building Open Knowledge Graph. The two studies do not evaluate GKG but evaluate the performance of the prototype they built.

CHAPTER 3

GKG AND SATORI ENTITY TYPES AND THEIR HIERARCHY

In this chapter, we aim to have a deeper understanding of the kinds of entities these semantic search engines are using and how they tend to organize and structure their data. This is a very important aspect of the semantic search engines and it will help to have a better understanding of the architecture of the semantic search engines and what they can accomplish with their data. The output of this chapter (the entity types indexed by the engines) will be used to investigate the availability of list search services provided by the semantic search engines.

3.1 INTRODUCTION

One major factor that will determine the functionality and usability of both Google's Knowledge Graph and Bing's Satori is the type of entities they have indexed and the number of entities in their database. However, building a large knowledge base that will scale semantic search engines to a new level will be a big challenge since most of the data on the web are unstructured even though webmasters are gradually implementing the schema.org [52] and other standards which add semantic to web pages. Having a lot of entities of different types is very important. Both GKG and Satori can only deal with information about entities they have indexed. The larger the amount of entities and relationship they have, the more they will potentially put a drag on the performance of search [67]. Both GKG and Satori are still in their infant stage because there's no way for search users to fully take advantage of the relationship mapping in these semantic data stores [67], but they are working to deliver a number of improvements and increment to their entity databases.

Another factor that will determine the usability of GKG and Satori is their data organization. Organization of knowledge is very important. It makes the data more organized, meaningful and easy to understand. By organizing classes into a hierarchy, one class (sub-class) can extend another class (super-class). This will also reduce storage space and other resources instead of building classes independently. Many types of ontologies and taxonomies have been developed to organize the human knowledge such as [10], [13], [12], [31]. The basic unit of knowledge representation is called concepts, classes or entity types. Concepts represent a class of entities. While restaurants is a concept, one specific instance of a restaurant is an entity. Concepts may have a hierarchical structure using isA relationships among them. Chinese restaurants, Turkish restaurants, Italian restaurants may all be a sub-concept of restaurants. Similarly, restaurants concept may be a sub-concept of a business concept. In addition, one concept may have more than one super concept. A Chinese restaurant may also be a sub-concept of historical place concept. Probase [10] and Cyc [31] used this kind of hierarchical organization of concepts. While Probase build the concepts algorithmically from a web page collection, Cyc knowledge base is constructed manually. Probase has more than 2 million concepts and Cyc has more than 500 thousand classes. Another type of concept organization is to have two levels of a hierarchy. Related concepts are grouped as a domain and each concept belongs to a single domain. For example, all commercial concepts can be grouped in a single domain, or all musical concepts can be grouped in a single music domain. However, concepts don't have an isA hierarchy among them. They are neither a sub-concept nor a super-concept of any other concept. Yahoo researchers proposed this kind of concept organization for web search in 2009 [19]. Freebase [12] also uses this kind of concept organization. Currently, Freebase has 76 domains and about 2000 concepts [12]. Although, concepts cannot have any relationships among them, instances of concepts may have many types of relationships. A professor entity may have hasA relationships with publication instances, worksIn relationship with a university department entity and hasBeenAwarded relationship with some award entities.

3.1.1 Research Questions for Chapter 3

In this chapter, we answer the first research question of the thesis with three sub questions:

1. What kinds of entity types GKG and Satori support?
2. What kinds of entity type organization do they use?
3. Do all entities of the same type have the same attributes?

3.2 METHODOLOGY

The method we used to investigate the coverage of classes in GKG and Satori is to use classes from an entity database and test the availability of their instances. There have been a lot of entity databases that are built to understand the real-world entities and their relationship. Among the prominent one include: Probase [10], Cyc [31], Freebase [12], Yago [13] and Dbpedia [14]. However, we use Freebase in our studies for the following reasons.

- Both GKG and Satori are using Freebase data to gather information about entities in their knowledge bases.
- Freebase data is freely available for research.
- It contains a lot of classes and entities and it is the largest open-source knowledge base.
- It has a good entity dataset

Freebase has over 39 million topics (entities) about real-world entities like people, places buildings, geographical features, things etc. Each topic or entity belongs to a class in Freebase and a class itself belongs to a certain domain. In Freebase, there are about 2000 classes and 76 domains [12]. Entity types in Freebase ranges from physical entities such as people and hotels; to Artistic or media creation such as movies and music, to classification such as noble gas, to Abstract concepts such as love, to School of thought such as impressionism and religion [64]. An entity in Freebase may be viewed from different perspectives for example; Barack Obama is a person, an author, and a politician. In order to

capture this multi-faceted of entities, Freebase introduce the concept of types and therefore, topics may have any number of types assigned to them.

Because of the large amount of classes and entities in Freebase, we do not test the availability all of the entities or classes in GKG and Satori, instead, we selected 100 of these classes and 10 entities (instances) from each of the 100 classes making 1000 entities. In selecting these classes, there are two issues we aimed at: first is to select these classes in such a way that we will have a large coverage of the classes, since some classes may be related or covered by other classes. For example the Author and Artist class whose instances are all people are said to be covered by the people class. Second, we avoid been bias to either GKG or Satori in selecting our classes.

3.2.1 Data Collection (Classes and Entities)

Each of Google Knowledge Graph and Bing's Satori has their own aims of building a web of data. Depending on this, each of them has some particular areas or classes for which it gives more priority in building its dataset. Therefore, we try to be fair in choosing our classes for the evaluation process. Also, we try to choose classes in such a way that they will cover large percentage of the knowledge data in both GKG and Satori.

The method of class selection is by randomly selecting the classes from a file we created that contains all the Freebase classes and domains. We adapt to this method because of the following reasons as stated above:

- (a) **To have a large coverage:** instead of selecting some few important domains to select our classes from, we randomly select the classes from all the Freebase available classes. Selecting these classes at random will help us have a large coverage since each class has equal probability of selection even though their domains have different probability since some domains have more classes than the others.
- (b) **To avoid being bias to either Google Knowledge Graph or Satori:** By selecting our class at random, we have avoided being bias to either Google Knowledge Graph or Satori.

For each of the class generated, we check on the fitness of its instances and decide/judge on whether it fit or not. By fitness here, we mean whether the class has meaningful instances about real world entities or not. For example, some classes may contain abstract instances, numeric values or relationship between other classes etc. These types of classes are not good for our evaluation research.

The classes we removed are due to one or more of the following conditions:

- **Few instances:** Classes that have few instances of less than 10 are likely to be irrelevant and less important. These kinds of classes may be given less priority when indexing classes in the search engines. An example of these kinds of classes include: Roller Coaster Train Configuration, Galactic interaction type, Event promoter, Unit of Conductivity, etc. which have less than 6 instances.
- **Classes covered by other classes:** In Freebase, an entity can belong to more than one class. This connects classes together to build a relationship. However, one class may cover the entities of another class. For example, the person class covers the following classes: Astronomer, Baseball Player, Baseball Manager, Animal owner, Ship owner, Comic Book Letterer, Computer Scientist, Film set decorator etc. these classes contains people as their instances which all belong to the person class. Similarly, the Region class is said to cover the following classes: Scottish council area, Chinese county, Mexican municipality, French Region, South Korean province, US Territory, Hong Kong district. The aim of removing classes that have been covered by other class is to have a large coverage of classes to be tested in the semantic search engines, Instead of testing a large pool of related or similar class.
- **It looks like a relationship:** Some class may be built by Freebase only to connect other classes, these classes may not have instances define. Example of such classes includes: Automotive: Option/Trim Relationship (Mediator type that links option, cost, MSRP and trim levels.), Organism Classification Placement (connects higher and lower organism classification with the authority/group/entity that places them in this relationship), Exchange rate (record historical rates of exchange between currencies).

- **It has irrelevant records:** Some classes may contain records that are irrelevant such as numeric values or statistical values. Example of these kinds of classes includes: Player Passing Statistics, Atomic mass, Camera Sensor Size, Mountain age etc.
- **Inaccessible instances:** some instance may not be accessible at the time we are gathering the classes and instances. As of January 2014, the following classes could not be access in Freebase: Chemistry: Radioactive decay mode, Exhibition: Exhibition venue, Measurement Unit: Dated Percentage, music: Musician, Organization: Nonprofit organization.
- **Have no distinguishing features:** Some classes may not have clear features or entities. The may contain instance that don't represent real-word entities. These classes are for additional information about entities. Example of these kinds of classes includes: American Football: Player Rushing Statistics, Biology: Plant Disease Conditions, Books: Book Binding, Organization: Organization termination type, Physical Geography: Mountain age, Rail: Steam locomotive wheel configuration etc.
- **Has compound value type:** Compound value types (CVTs) are used in Freebase to represent complex data, where each entry consists of multiple fields. They are used to accurately model complex relationships between topics. Example of these kinds of classes includes: Biology: Breed registration, Government: Legislative committee membership, Measurement Unit : Money value, music: Musical Group Membership, Opera: Opera character voice, Religion : Religious Organization Leadership, Sports : Sports League Draft Pick, Transportation: Transit System Length, TV: TV Producer term etc.
- **Have no instances:** some class may have no instances defined. As of January 2014, the following classes has no instances in Freebase: Government: Polled area, Measurement Unit: Dated BTU

When selecting our classes, we ensure that no class is selected more than once. For the case of selecting our instances (entities), we selected the instances in sequence from the

first instance to the tenth instance in that order as they appear in the Freebase page. But there is an exception to this order, which is the case when instances look similar and there are other different instances available. For example, the following instances appear sequentially in the “Film festival event” class: 2005 Cannes Film Festival, 2007 Cannes Film Festival, and 2006 Cannes Film Festival. In this case, we therefore, try to minimize this kind of issues by selecting only one or two of this kind of instances to provide a chance for checking other kind of instances in a class.

3.2.2 Search Engine Settings

In order to be consistent in testing the queries in both GKG and Satori, we set both engines to their US interfaces throughout our research. We used *www.google.com* for GKG and *www.bing.com* for Satori. In the case of Satori, we also changed the location to “United State-English” from the settings in Bing’s home page. Users can change their location by clicking the settings icon by the top right corner of the search engine’s home page. Inside the settings, the location can be changed by clicking the country/region link and then selecting United State-English. This is very important because users may not get result without the settings especially in the case of Bing. Also GKG or Satori may not be available in some countries, as they both reported to have started from some few countries. Moreover, we predict that the location used in searching can also affect the quality of the result obtained in some semantic search engines.

3.2.3 Testing the Entities

For each of the valid class we selected, we check for the availability of 10 of its instance in both Google Knowledge Graph and Satori. We therefore checked for the availability of 1000 entities. We marked a class as present in either GKG or Satori if at least one of its instances is available in either GKG or Satori respectively; we marked a class as absent in either GKG or Satori otherwise. For each of the entity present in Google Knowledge Graph or Satori, we also examine the information displayed on the SERP. The following are the information we examine about the entities.

- Availability of entities of a class in Google Knowledge Graph and Satori

- The class label for that entity in both Google Knowledge Graph and Satori.
- Hierarchies in the information displayed for entities.

3.3 EXPERIMENTAL RESULTS

The result we obtained after testing 1000 instances from 100 classes is as shown in Figure 3.1. From the figure, it shows that 40 out of 100 of the classes are not in Google Knowledge Graph, in other words, 60% of the classes tested are available in Google Knowledge Graph. Similarly, 34 of the classes tested are not in Satori which means that 66% of the classes appear to be in Satori. Also 75% of the classes were found to be either in Google Knowledge Graph OR Satori. The OR operator is similar to binary “OR” operator and it is true whenever a class is present in one or both of the semantic search engines. Lastly only 51% of the classes are present in both Google Knowledge Graph “AND” Satori. The “AND” here is also like a binary AND operator and it is true only if a class is present in both Google Knowledge Graph and Satori.

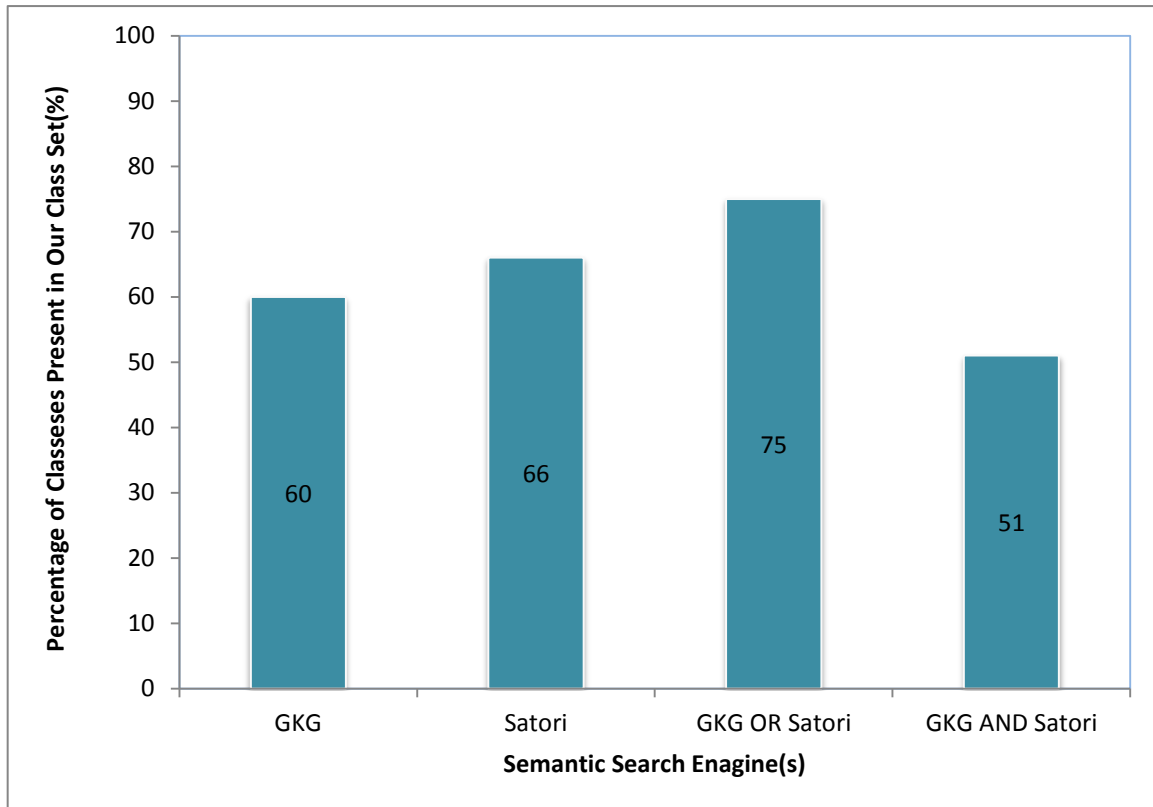


Figure 3.1: percentage of classes presents in GKG and Satori from our class set

It is important to note that the above result does not say anything about the class coverage in either Google Knowledge Graph or Satori, it only gives some of the classes that are present or absent in our list of classes. This research has therefore not included class coverage. We hope to include class coverage in our future work. In the class coverage, we will investigate how much of the entities of a class are recognized or present in either of the semantic search engines we used in this research. For example, how much of people, universities, celebrities, buildings or geographical features are recognized in either Google Knowledge Graph or Satori.

Even though we have not investigated all the classes in either GKG or Satori, and neither of these technologies has explicitly specified the type of classes in their dataset. Below are the lists of some classes covered by these technologies base on the research we conducted.

- a. **Classes Covered by GKG:** The classes that are covered in GKG include: Amusement Ride, Roller Coaster, Houses, Buildings, Offices, Museums, Planets, lakes, rivers, mountains, valleys, waterfalls, Constellations, Books, Films/movies, Musical Albums, Animals, Companies, Job Titles, Shopping Centers and Malls, Comic Strips, Academic Institutions, Disasters, Recurring Events, Fictional Object Destruction Methods, Film Subjects, Beer, Beverage Types, Distilled Spirit Types, Countries, Regions, Drugs, Drug Classes, People (Actors, politicians, Artists, Lobbyist etc.), Tribes, Religious Jurisdiction, Theater, Tourist Attractions, Transportation Mode, Hotels, Sport Teams, Animal Breeds, Gene, Hybrids, Consumer Products, Engines, Isotopes, Fictional Characters, Fictional Universes, Food, Restaurant, Hormones, Hospitals, Organization Committees, Railway Types, Transportation Companies and Automobiles.
- b. **Classes Covered by Satori:** the following are the classes we found to be covered by Satori: Amusement Ride, Lighthouse Construction Materials, Roller Coaster, Houses, Planets, Constellations, Near-Earth Objects, Novels, Movies, Books, Musical Albums, Recurring Competitions, Baseball Divisions, Sport Teams, Animals, Venture Investors, Shopping Centers, Comic Strips, Conference Subjects, Cricket Leagues, Academic Institutions, Disasters, Recurring Events, Fictional Objects Destruction Methods, Film Subjects, Film Festivals Events, Beer, Foods, Beverage Types, Infused Spirits, Playing Card Games, Connection Protocols, Countries, Regions, Drugs, Drug Classes, Musical Groups, Musical Performance Role, Musical Instruments, Olympic Mascot, People, (Actors, politicians, Artists, Lobbyist etc.), Religious Jurisdiction, Order of Chivalry, Theater, Tourist Attractions, Transportation Mode, TV Seasons, Visual Art Forms, Baseball Leagues, Animal Breeds, Hybrids, Consumer Products, Isotopes, Fictional Characters, Fictional Universe, Restaurants, Geological Formations, Measuring Instruments, Hormones, Hospitals, Medical Specialties, Organization Committees, Musical Recoding, Transport Operators, Digital Cameras and Roads
- c. **Classes Not covered by GKG:** The following classes are not covered by GKG: American football game, Lighthouse construction material, Competition type, Recurring competition, Baseball Division, Cricket League, Student radio station,

Degree, Fictional Plant, Film festival event, Playing card game, Internet Protocol, Patent office, Infectious Disease, Bones, Manufactured drug form, Nerve, Olympic mascot, Olympic event competition, Olympic games, Fundamental interaction, Religious Jurisdiction Category, Religious order, Football world cup, Calendar System, TV Season, Art Period/Movement, Visual Art Form, Legislative committee, Measuring Instrument, Medical specialty, Vein, Satellite Type and Video Game Rating System.

- d. **Classes Not Covered by Satori:** The following classes were not covered by Satori: American football game, Type of planetographic feature, Job title, Student radio station, Degree, Fictional Plant, Patent office, Infectious Disease, Bone, Manufactured drug form, Nerve, Olympic event competition, Olympic games, American Indian group, Fundamental interaction, Religious Jurisdiction Category, Religious order, Football world cup, Calendar System, Hotel Grading Authority, Art Period/Movement, Gene, Student organization, Engine, Legislative committee, Vein, Railway type, Satellite Type, Video Game Rating System, Unit of Data Transmission Rate, Locomotive class, Automobile generation, File Format and Anatomical structure.

The above classes in [c and d] may not have been covered by GKG and Satori probably because they are yet to be implemented by these technologies or they feel they are less important or irrelevant to their aims of building web of data

- e. **Classes covered by GKG that are not covered in Satori:** the following classes are covered by GKG but not in Satori: Type of planetographic feature, Job title, American Indian group, Hotel Grading Authority, Gene, Student organization, Engine, Railway type and Automobile generation
- f. **Classes covered in Satori that are not covered in GKG:** the following classes are covered by Satori but not in GKG: Lighthouse construction material, Competition type, Recurring competition, Baseball Division, Cricket League, Film festival event, Playing card game, Internet Protocol, Olympic mascot, TV Season, Visual Art Form, Measuring Instrument, Medical specialty, Musical Recording and Digital Camera.

From our observations, the kinds of classes covered by these semantic engines are classes that are very popular and interacted by people in their daily activities. Most classes covered are what people are mostly interested on the web such as celebrities, tourist attractions movies, songs, restaurants, hotels, food items, geographical features (waterfalls, rivers, oceans, lakes), and many things that bring fun and interest people.

An investigation of the classes not indexed by both search engines shows that most of the classes are usually not as important as the ones indexed. Some of these classes may include: File Format, Unit of Data Transmission Rate, Calendar System, Student radio station, Degrees, Fictional Plants, Infectious Diseases, Bones, Olympic Games, satellite types etc. These classes may not be that much interested by web users. The search engines may probably be using data from their query log to monitor the kind of thing users are after. Among these classes, the Medicine domain has the largest turnout with six (6) of its classes not indexed by the search engines which includes: Infectious Diseases, Bones, Manufactured drug form, Nerves, Veins and Anatomical structures.

In general, Satori has more classes in their engine than GKG having 66% of the classes of the classes we tested, while GKG has 60% as shown in Figure 3.1. Since the instances of the class tested are all entities, Satori could perform better for searches about single entities. However, having lots of classes is one case and utilizing those classes to efficiently answer users demand is another thing. As in the case of the traditional keywords search engines, indexing lot of documents is important but the search engine that indexed the most data may be the worse engine if they could not use the advantage of the data or have a good algorithms that will return the most relevant documents.

Satori may include more entities or company products to aid transactions, while GKG is trying to cover more entities to aid informative queries. Products in Satori not in GKG include: Lighthouse construction materials such as Brick, Concrete, Cement , Granite ,Limestone etc., Playing card games such as Blackjack, Contract bridge, Go Fish, Poker, Pinochle, Sheepshead etc., Digital cameras such as Canon PowerShot A75, Nikon D1, Nikon D80, Leica Digilux 2, Kodak DX7590, Nikon Coolpix S1etc., and Measuring

Instruments such as Anemometer, Atomic clock etc. An investigation of the classes in GKG that are not in Satori shows that most of these classes are not product like but are more like informative classes. These classes includes: Type of planetographic features, Job title, Student organizations, Hotel Grading Authorities, Railway types etc.

An investigation of the entity types not found in GKG and Satori shows that a good number of these entity types are events centric i.e. they are having events as there instances such as American Football Game, Recurring competition, Football World cup, Olympic Event Competition, Film Festival etc. 32.5% of the classes not found in GKG are these type, but the value decreases in Satori by almost a half, which means that Satori have indexed more of these classes in its entity dataset than GKG with only 17.6% of the classes not found in it. We think this type of entity types were not found much in GKG and Satori probably because they are less searched by web users or Google and Bing indexes entity types given priority to some set of entity types.

3.4 HIERARCHICAL STRUCTURE OF ENTITY TYPES

The main task of semantic web is to organized web data into a more meaningful form. This can be achieved by defining entities from the web data, linking them and classifying them into types. In addition, hierarchies may be defined to further organize and structure the data in a more formal and comprehensive form. [65] Proposed a 3-level hierarchy for Named Entity types, while [66] introduce the task of hierarchical target type identification, which uses ranking mechanism to identify the type of relevant results with respect to a given ontology from a query. Hierarchy among entity types forms a logical connection between them, and helps in making the entity dataset more meaningful. A hierarchy in the entity type can help users to find interesting things about what they are looking for and even navigate to find more information about related concepts. For example, for a query about a person such as “Barack Obama”, one should be able to find information about his basic bio-data and even go deeper or broader to find information about his parents, grandparents, children, siblings, etc. However, building hierarchy among entity types is challenging and not an easy task. Nilesh Dalvi et.al in their paper [19] point

out that there are representational, expressiveness and computational cost issues to be considered when representing or building hierarchies among entity types and entities. Beyond these issues, there are also issues of feasibility of extracting information from the data set and also the ability to reliably interpret and match users query to what is available in the entity database. Clearly, search engine vendors would represent and structure their data set based on what is efficient to them and would also satisfy their users.

Part of our research is to discover whether or not GKG and Satori have implemented hierarchies among their entity types. Our investigation has shown that both GKG and Satori may not show taxonomical hierarchies of entity types to their users, based on what they displayed in their SERP. We believe that they might have no hierarchies defined internally in their entity dataset. They may be similar to Freebase in which there is no hierarchy among entity types. Only some related concepts are grouped in the same domain.

In both GKG and Satori, hierarchies are shown in very few classes such as animal, drugs and the person class. But these hierarchies are not based on the taxonomical scenario, since there is no kind of inheritance such that a class can inherit all of the features from its superclass. The hierarchies in this case are implemented as features and there is no is-a relationship between entity types but users can view and navigate to higher or lower classification.

3.4.1 Hierarchy in GKG

In GKG, the hierarchy can be seen in the animal class. In this case, an animal may have a lower and higher classification; which allows users to view higher and lower classification of the animal at the feature level. An example of this classification can be seen for a search of “fish” in Google as shown on the right hand corner of Figure 3.2. Even in the animal class, this classification does not appear in some instances of the class. For example, as of November 2013 a search for cat or dog in Google does not show the higher and lower classification as shown in Figure 3.3.

The image shows a Google search result for the query "fish animal". The search bar at the top contains the text "fish animal" and a microphone icon. Below the search bar, there are navigation tabs for "Web", "Images", "Maps", "Shopping", "News", "More", and "Search tools". The search results show "About 328,000,000 results (0.26 seconds)".

On the left side, there are several search results:

- "Images for fish animal - Report images" with a row of four fish images.
- "Fish - Wikipedia, the free encyclopedia" with a link to en.wikipedia.org/wiki/Fish and a brief definition: "A fish is any member of a paraphyletic group of organisms that consist of all gill-bearing aquatic craniate animals that lack limbs with digits. Included in this ...".
- "BBC Nature - Fish" with a link to www.bbc.co.uk/1/1/2009/09/090928_fish.shtml and a snippet: "Fish. It's strange but true that there's really no such thing as fish! Unlike with mammals and birds, not all the animals we call fish - aquatic, vertebrate animals ...".
- "Strange Animals: Strange Fish - YouTube" with a link to www.youtube.com/watch?v=kZuVyPRhFME and a snippet: "28 Sep 2009 - Uploaded by AutoCAD Tutorial An oarfish can easily be mistaken for a sea serpent, but they are really the longest bony fish in the sea ...".

On the right side, there is a knowledge panel titled "Fish". It includes:

- A small image of a pufferfish.
- The word "Fish" in large font, followed by "Animal" in smaller font.
- A definition: "A fish is any member of a paraphyletic group of organisms that consist of all gill-bearing aquatic craniate animals that lack limbs with digits." with a "Wikipedia" link.
- Classification details:
 - Rank: Species
 - Mass: 2.3 – 2.7 kg (Oreochromis aureus, Adult)
 - Higher classification: Gnathostomata
 - Clutch size: 160 – 1,600 (Oreochromis aureus, Female)
 - Length: 35 cm (Oreochromis mossambicus, Adult), 13 – 20 cm (Oreochromis aureus, Adult)
 - Lower classifications: Osteichthyes, Chondrichthyes, Acanthodii, Placodermi
- A "Feedback/More info" link at the bottom right.

Figure 3.2: GKG result showing higher and lower classification

Also, in person class, hierarchies are shown for some entities by displaying the parent and children of that person. It shows that the hierarchy in the person class is not as strong as the one shown in the animal class, probably because more animals are covered than people in GKG entity database. One may want to reach person's A grandfather starting from that person, but on reaching person's A father say 'B', one may get halt probably because B's father, also A's grandfather is not indexed in the entity dataset, or users may not be interested in B's father and therefore may not be displayed as B's feature. For example, we would like to know Barak Obama's grandfather; a search for "Barak Obama" has his father i.e. Barack Obama Sr. on the result page as one of the features probably because people are interested in knowing Barak Obama's father. Clicking on Barack Obama Sr. takes us to his page and luckily his father Hussein Onyango Obama is also one of the features. But on clicking Hussein Onyango Obama, no information is seen for Hussein Onyango Obama who is the grandfather of Barack Obama.

Google cat

Web Images Maps Shopping News More Search tools

About 457,000,000 results (0.41 seconds)

Cookies help us deliver our services. By using our services, you agree to our use of cookies. [OK](#) [Learn more](#)

[Cat - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Cat
The domestic cat (*Felis catus* or *Felis silvestris catus*) is a small, usually furry, domesticated, and carnivorous mammal. It is often called the housecat when kept ...
[African wildcat - Creme Puff - List of cat breeds - Cats and humans](#)

[Centre for Alternative Technology](#)
www.cat.org.uk/
CAT aims to empower people to live a more sustainable life. Through a combination of post graduate, short, and school courses and practical onsite examples.

[News for cat](#)

[Under the Paw](#)
Book cover image

[Don't judge a cat book by its cover](#)
The Guardian (blog) - 4 hours ago
Publishers will tell you that a cute animal on the cover is what sells in the supermarkets, but my readers' responses suggest otherwise, writes ...

[Dear Prudence \(the Advice Cat\): Advice for Cat Lovers From a Cat's Point of ...](#)
Huffington Post - 1 hour ago

[Ancient Chinese cat bones shake up domestication theory](#)
The Guardian - 2 days ago

Cat
Animal

The domestic cat is a small, usually furry, domesticated, and carnivorous mammal. It is often called the housecat when kept as an indoor pet, or simply the cat when there is no need to distinguish it from other felids and felines. Wikipedia

Scientific name: *Felis catus*
Lifespan: 12 – 14 y (Male, Domesticated, Newborn)
Gestation period: 64 – 67 d
Mass: 4 – 5 kg (Domesticated)
Daily sleep: 12 – 16 h
Rank: Species

Breeds

Figure 3.3: GKG result not showing lower and higher classification for searching cat.

3.4.2 Hierarchy in Satori

In Satori, hierarchy in the animal class differs a little bit from GKG. Satori gives the biological classification of the animal such as species, phylum, order, class, family to classify each animal. They also list the animal's biological class in the 'consist of' property. Sometimes, they append the prefix 'sub' or 'super' to the biological classification such as sub-phylum or super-order to further classify animals. In addition, they state the class each animal belongs to in the 'belongs to' property which is similar to the higher classification in the case of GKG as shown in Figure 3.4. Similar to GKG, Satori also has hierarchy defined in the person class, in which people have parent and children as features.

WEB IMAGES VIDEOS MAPS NEWS MORE 5 of 5 Farouk Mus...

bing panthera genus

188,000 RESULTS Any time ▾

[Panthera - Wikipedia, the free encyclopedia](#)
 en.wikipedia.org/wiki/Panthera ▾
Panthera is a **genus** within the Felidae family that was named and first described by the German naturalist Oken in 1816. The British taxonomist Pocock ...
 Name · Characteristics · Evolution · Classification

[Genus panthera | Define Genus panthera at Dictionary.com](#)
 dictionary.reference.com/browse/genus%20panthera ▾
genus panthera noun lions; leopards; snow leopards; jaguars; tigers; cheetahs; saber-toothed tigers [syn: **Panthera**]

Related searches for **panthera genus**
[Animals in the Panthera Genus](#) [Panthera Species](#)
[Black Panther Genus and Species](#) [What Is Panthera](#)
[What Does Panthera Tigris Mean](#) [Panthera Leo](#)

[Panthera | Define Panthera at Dictionary.com](#)
 dictionary.reference.com/browse/panthera ▾
 noun . a **genus** of chiefly large cats that includes the snow leopard, tiger, leopard, jaguar, and lion, most having the ability to roar.

[Tiger - Wikipedia, the free encyclopedia](#)

Panthera
 Panthera is a genus within the Felidae family that was named and first described by the German naturalist Oken in 1816. The British taxonomist Pocock revised the classification of this genus in 1916 as comprising the species tiger, lion, jag... +
 en.wikipedia.org
 www.animalpicture...
 Scientific name: Panthera
 Biological classification: Genus
 Consists of: Jaguar · Lion · Leopard · Tiger · European jaguar · Tuscan lion · Panthera youngi · Panthera palaeosinensis +
 Belongs to: Pantherinae

People also search for






 Cougar  Snow leopard  Puma  Carnivora  Lynx

Figure 3.4: searching for panthera in Satori

However, this type of entity classification is only shown among the entities of few entity types in both Google knowledge graph and Satori. The semantic engines do not allow their users to surf deeper and wider into their entity database in other classes. They do not organize most of their data in hierarchy. For example all geographical location in the world can be organized into a hierarch in which one region or location falls into another. For example, East Los Angeles is in Los Angeles, Los Angeles is in California, California is in Southwestern United States and Southwestern United States is in USA. By searching one of the locations, one should be able to navigate to lower or higher region in the hierarchy. Organizing data in this way can give users the advantage to surf the entity database and discover related and interesting information about places or things.

3.5 FEATURES OF ENTITIES DISPLAYED IN GKG AND SATORI

An entity type consists of group(s) of entities having similar properties .The entity type of an entity determines the kind of properties that entity would have. For example, in the person class/type, a person may be defined by some of the following properties: a name, nick name, place of birth/nationality, date of birth, parent, children, occupation, area of

specially, Education/school attended, books, awards etc. It is obvious that an entity may or may not have values for some of the properties defined by its class. In particular, the properties are what uniquely define an entity of a class. The properties are what GKG and Satori display in response to users query for an entity. Both GKG and Satori do not display all of these properties, the properties displayed by these semantic search engines do not only depends on the entity type, but it also strongly depends on the entity itself. One question to ask is; are entities of the same class having the same features?

We find out that different properties are displayed for entities of the same class, even though they may have those properties defined. For example, different properties may be returned for two different scientists or professors. To be more specific, consider two different computer scientists such as Tim Berners-Lee and Vint Cerf. The properties shown for a search of “Tim Berners-Lee” in GKG are: born, books, awards, education, nationality and parent, while only born, award and education properties are shown for a search of “Vint Cerf”. The properties displayed by GKG and Satori may depend on some ranking algorithm known to them. [68] Reported that GKG uses their query log to determine the kind of properties users are interested for a particular entity. For example, the nationality and parent properties displayed for a search of “Tim Berners-Lee” in GKG might be as a result of users’ interest and curiosity for those information. Notice that Vint Cerf may also have those properties defined but probably users are not much interested to know about those information. [69] Proposed a method that allows contrasting of class definitions found in Semantic Web vocabularies with the attributes of objects that users are interested in.

3.6 CONCLUSION

In this research, we investigated the kinds of entity types in GKG and Satori. From our observations, the kinds of classes covered by these semantic engines are classes that are very popular and interacted by people in their daily activities. Most classes covered are what people are mostly interested on the web such as celebrities, tourist attractions, movies, songs, restaurants, hotels, food items, geographical features (waterfalls, rivers, oceans,

lakes), and many things that bring fun and interest people. An investigation of the classes not indexed by both search engines shows that most of the classes are usually not as important as the ones indexed. Some of these classes may include: File Format, Unit of Data Transmission Rate, Calendar System, Student radio station, Degrees, Fictional Plants, Infectious Diseases, Bones, Olympic Games, satellite types etc.

Our results have shown that not all the classes in Freebase are indexed in either GKG or Satori. Based on the research we conducted, 60% of the classes we used from Freebase are available in GKG and 66% of these classes are available in Satori. Similarly, 75% of the classes were available in either GKG or Satori. Lastly 51% of the classes appear to be available in both GKG and Satori.

An investigation of the entity types not found in GKG and Satori shows that a good number of these entity types are events centric i.e. they are having events as there instances such as American Football Game, Recurring competition, Football World cup, Olympic Event Competition, Film Festival etc. 32.5% of the classes not found in GKG are these type, but the value decreases in Satori by almost a half, which means that Satori have indexed more of these classes in its entity dataset than GKG with only 17.6% of the classes not found in it. We think this type of entity types were not found much in GKG and Satori probably because they are less searched by web users or Google and Bing indexes entity types given priority to some set of entity types.

Our investigation has also shown that GKG and Satori may have fewer classes than some of the entity dataset like Probase, Yago, and Cyc which have thousands to millions of entity types. The reason behind this may be as a result of human intervention in selecting what kind of entity types to be included or not included in their dataset rather than having a particular algorithm that automatically find and index new entity types in the dataset, regardless of what type it is. The human intervention is aimed at providing a high accuracy to satisfy users with the appropriate answers for their query. If entity types are to be extracted and indexed algorithmically or automatically, as it is done in most document search engines, GKG and Satori will have a large amount of entity types in their dataset,

but may tend to be noisy given the vastly different content in the large collection of sites [19]. Therefore, the information stored in this case may not meet the user's demand.

Our investigation has also shown that both GKG and Satori do not show taxonomical hierarchies of entity types to their users, based on what they displayed in their SERP. We believe that they might have no hierarchies defined internally in their entity dataset. They may be similar to Freebase in which there is no hierarchy among entity types. Only some related concepts are grouped in the same domain.

We also observed that different properties are displayed for entities of the same class, even though they may have those properties defined. The search engines only try to return information that their users are interested in and not all information about the entity they searched.

CHAPTER 4

LIST SEARCH SERVICES FOR ENTITIES

4.1 INTRODUCTION

Both Google and Bing have been working tirelessly in providing users results without clicking on the results in their regular search engine results page (SERP). They tend to answer most of the user's queries about famous people, places and things. Although most of the queries users asked about entity searches are single entity seeking queries, there are a lot of cases when users may inquire for a list of entities from the search engines. GKG and Satori provide a means of displaying a list of entities in a carousel interface for queries about list of entities. The carousel was shown initially when a user clicks on the 'people also search for' link from a search about an entity [70]. Both GKG and Satori usually give five entities related to the searched entity in addition to the information about the searched entity. However, users can expand it when they click on the 'people also search for' link. The related entities will then be displayed in the carousel.

Currently, the carousel result page has been extended to new queries. For some queries that request a list of entities, the carousel result page is shown with the list of entities. For example, the carousel is shown for the search queries about multiple entities like 'universities in Canada', '2013 movies', 'books by Stephen King' etc. For the query 'universities in Canada', a list of top universities in Canada is returned in the carousel. This requires the semantic search engine to have the list of universities in Canada and rank them. Only top ranking universities are returned in the carousel since they have a limited space.

This makes the navigation of entities much easier to users. Users can navigate through many universities without leaving the same search engine result page. Providing the list of entities for user queries is a very important feature of semantic search engines.

Displaying list of entities has been of utmost important in helping the user's find information about the real-world entities and how they are connected. The carousel result displayed at the top of the SERP has been taking much of user's attention with its image interface. [71] Reported that "these types of results will "disrupt" traditional searcher behaviors that heavily favored the first search results by spreading clicks more evenly across the page. With a horizontal display, it might be equally advantageous to be third as it is to be first in order." They seem to be showing the carousel result more frequently for user's queries and they have been updating the looks of the carousel results [72], [73]. The carousel result sometimes includes dropdown boxes to help users filter and find the information of their interest. However, from our result, it shows that both search engines are still in their infant stage as majority of their classes do not still produce the carousel for searches about multiple entities. The classes that produce the carousel are those that are of more interest to their users like music, movies, tourist attractions, rivers, mountains, lakes, academic institutions etc. which means that they may probably include the carousel result for other classes with less priority.

Both GKG and Satori are expanding their carousel to include more classes. GKG have expanded their carousel results in addition to the restaurant and hotels searches to include more topics [73]. Building this browse-able interface requires a deep understanding of the entities and the relationship between them. The entities shown on the carousel are based on a huge experience gathered from the entity databases.

4.1.1 Research Question

In this chapter, we investigate the extent of classes that are supported by list search through the carousel result page:

1. For what kinds of entity types, a list of entities is returned by semantic search engines in carousel result page? How common is it for semantic search engines to provide the list search service for entity types?

4.1.2 Multiple Entity Search Result

Unlike the usual ranked links returned by traditional search engines for user's queries, the type of results obtained by some semantic search engines strongly depends on the type of query or the intent of the user. One of the most interesting types of search involves a search for multiple entities. In fact searching about multiple entities has been very important by exposing how well the search engines understand the real-world. The entities returned by this kind of searches are connected by some certain attributes or constrains. For a search like 'rivers in London' the search engine should be able to understand that the primary entity is river and only search in the river domain or class but return only those found in London by scanning the attributes of each river. It is also possible to tag queries to results stored as cache so that they are easily return when users search for the same thing in order to save time.

GKG and Satori return a list of entities for searches about multiple entities or group of entities. The entities are returned in a horizontal rectangle interface at the top of the regular search engine result page. This horizontal list of entities is called '*carousel*'. GKG refer to this interface as 'knowledge graph carousel'. The entities on the carousel are browse-able left to right or right to left by clicking on the arrows by the sides of the carousel. As the entities are arranged in some order, we expect that the entities themselves are ranked using some mechanism by the search engines. Figure 4.1 shows a snapshot for a search of 'tourist attractions in Turkey'.

The image shows a Google search results page for the query "tourist attractions in turkey". At the top, there is a search bar with the query and a "Sign in" button. Below the search bar, there are navigation links for "Web", "Maps", "Images", "News", "Shopping", and "More". A "Search tools" button is also present. The main content area features a "Turkey > Points of interest" section with a carousel of 14 images and labels for various landmarks: Hagia Sophia, Sultan Ahmed Mosque, Topkapı Palace, Grand Bazaar, Basilica Cistern, Pamukkale, Chora Church, Süleymaniye Mosque, Istanbul Archaeology Museums, Golden Horn, Galata Tower, Temple of Artemis, Mount Nemrut, and Dolmabahçe Palace. Below this carousel, there are several organic search results. The first result is "10 Top Tourist Attractions in Turkey | Touroplia" with a snippet mentioning "Aug 15, 2013 - With so many amazing destinations a top 10 is bound to leave some great tourist attractions in Turkey out. So consider this list of destinations ...". The second result is "Images for tourist attractions in turkey" with a "Report images" link and a small image gallery. The third result is "Top 10 Tourist Attractions in Turkey - Historivius" with a snippet stating "Probably the most famous tourist attraction in Turkey, the Hagia Sophia is one of the ... the Basilica Cistern rightly ranks among Turkey's top tourist attractions." The fourth result is "Turkey Travel | Places to visit in Turkey | Rough Guides" with a snippet mentioning "a comprehensive list of the most and the best things to see in Turkey". On the right side of the page, there is a knowledge panel for "Turkey". It includes the Turkish flag, a map of Turkey, and the following information: "Country", "Turkey, officially the Republic of Turkey, is a contiguous transcontinental country, located mostly on Anatolia in Western Asia, and on East Thrace in Southeastern Europe. Wikipedia", "Capital: Ankara", "Dialing code: +90", "Currency: Turkish lira", "Prime minister: Recep Tayyip Erdoğan", "Official language: Turkish Language", "Government: Unitary state, Parliamentary republic", and "Destinations".

Figure 4.1: Example of list search results in GKG.

From Figure 4.1, the list of entities that matched tourist attractions in Turkey is shown at the top of the SERP in the black background. The entity ‘turkey’ is also recognized by the search engine as part of the result for the query as shown in the knowledge graph result space at the right hand side of the SERP. While it seems to be clear why they display ‘Turkey’ in the result, been the most pronouns entity in the query. However, GKG those not display the Turkey entity in the knowledge graph result page for the query ‘rivers in Turkey’, instead it display the ‘rivers’ entity (Figure 4.2). In contrast, Satori outputted the carousel and ‘Turkey’ entity for both queries. Figure 4.3 shows the result for searching ‘rivers in Turkey’ in Satori.

The screenshot shows a Google search for "rivers in turkey". The search bar is at the top with the Google logo. Below the search bar, there are navigation tabs for Web, Maps, Images, News, Shopping, and More. The main content area features a carousel of river images with labels: Tigris, Euphrates, Büyük Menderes River, Çoruh River, Göksu, Seyhan River, Aras River, Orontes River, Evros River, Gediz River, Yeşilirmak River, and Kızılırmak River. Below the carousel, there are search results for Wikipedia and a category page for "Rivers of Turkey". On the right side, there is a map showing a location in Ankara, Turkey, with the title "Rivers" and address "1946 Sk No:13, Çayolu/Ümitköy/Ankara, Turkey".

Figure 4.2: searching for 'tourist attractions in turkey' in GKG.

There is no standard or format for displaying results in the semantic search engines. Even for the multiple entity searches, there can be different types of carousel results depending on the type of queries. In the next paragraph, we identify three (2) different types of carousel outputs.

The screenshot shows a Bing search for "rivers in turkey". The search bar is at the top with the Bing logo. Below the search bar, there are navigation tabs for WEB, IMAGES, VIDEOS, MAPS, NEWS, and MORE. The main content area features a carousel of river images with labels: Euphrates, Tigris, Khabur River, Orontes River, Aras River, Pactolus, Kura River, Evros River, Kızılırmak River, and Büyük Menderes River. Below the carousel, there are search results for Wikipedia and a category page for "Rivers of Turkey". On the right side, there is a map showing a location in Ankara, Turkey, with the title "Rivers" and address "1946 Sk No:13, Çayolu/Ümitköy/Ankara, Turkey".

Figure 4.3: Example of list search results in Satori

4.1.2.1 Simple carousel result

This is the most common type of carousel output. They consist of horizontal entities that can be navigated by users. The operations that can be performed on this type of carousel is to click on individual entities to view its information on the knowledge graph or Satori result page or navigate to view the list of entities on the carousel. GKG return at most 51 entities in the carousel and about 12 entities per page. While Satori returns at most 50 entities in their carousel result and a maximum of 10 entities per page. Figure 4.1, 4.2 and 4.3 shows a sample of these types of carousel results.

In addition to the list of entities and the navigation icon, they may contain filters in the form of dropdown box that users can select to filter the entities. The dropdown boxes are shown on the top right of the carousel. There are usually one or two drop-down boxes. Example of this type of carousel appears for a search of ‘2012 movies’ in Google (Figure 4.4).

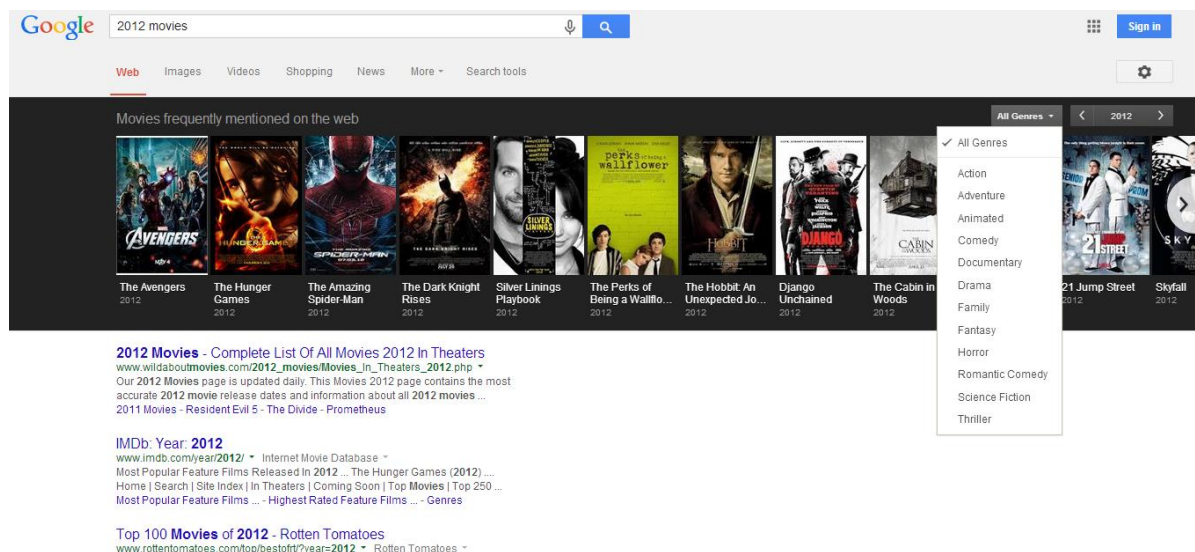


Figure 4.4: Carousel with double filters.

As shown in Figure 4.4, there are two filters in this case: the category of the movie (genres) and the year of the movie (date). The default category is ‘All genres’ but other options exist such as actions, adventures, animation, comedy, drama etc. However, users need to specify the year for the carousel to appear. Because searching for just ‘movies’ does not give any result in both GKG and Satori. The year specified in the query becomes the default date/year, but users can also select/filter by year from the carousel to that which they desire. The two filters are mutually exclusive, and the results in the carousel as well as the query are updated each time a user select an option from one of the filters. For example, while the year is still at 2012, selecting ‘action’ in the category/genre updated the query to ‘list of 2012 action films’. Of course the entities on the carousel are updated too to only action movies in 2012 year.

Another example is by searching ‘books by James Patterson’ (Figure 4.5). In this case, there is only one filter or drop-down box. There are three options, the default being ‘most popular first’ and the other options are ‘newest first’ and ‘oldest first’.

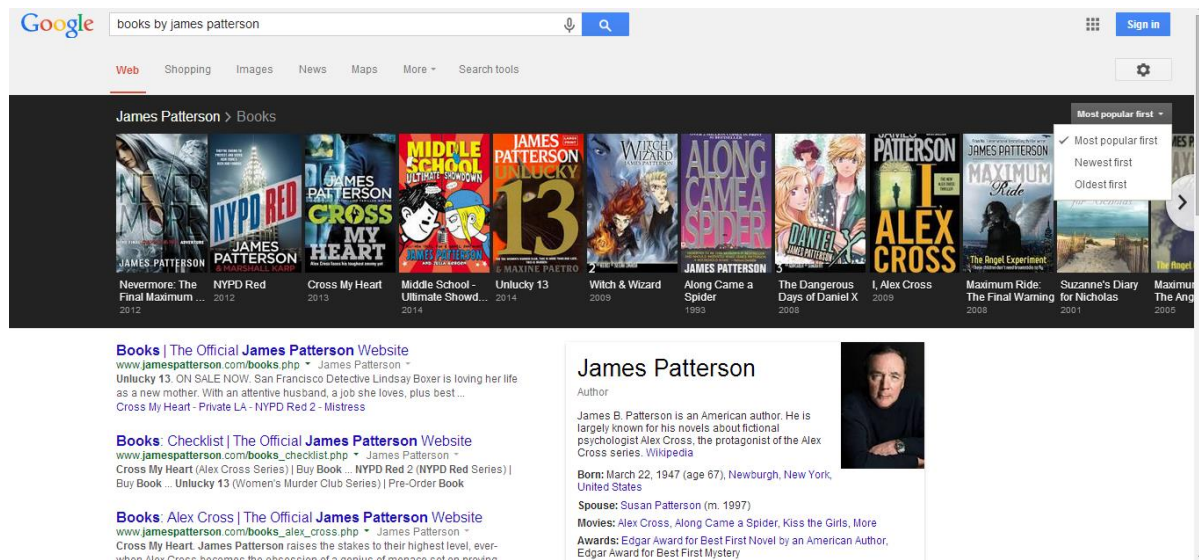


Figure 4.5: carousel result with single filter in GKG.

4.1.2.2 Complex carousel

This type of carousel is shown when users search for songs by an artist. It differs from the simple carousel. Unlike the arrangement of the other types of carousels that has only one horizontal list of entities, the entities in the complex carousels are arranged in phase or steps. Moreover, the icons representing an entity in the complex carousel does not look like an image. Figure 4.6 shows the complex carousel for a search of ‘songs by Michael Jackson’ in GKG.

The screenshot shows a Google search for "songs by michael jackson". The search bar is at the top with the text "songs by michael jackson". Below the search bar, there are navigation tabs for "Web", "Videos", "News", "Shopping", "Images", and "More". The main content area is titled "Michael Jackson > Songs" and displays a grid of song titles and durations. Below this, there is a "News for songs by michael jackson" section with several news snippets. To the right, there is a knowledge panel for "Michael Jackson" with a portrait and biographical details.

Song Title	Duration	Song Title	Duration	Song Title	Duration	Song Title	Duration
Love Never Felt So Good	3:55	Beat It Thriller • 1982	3:49	They Don't Care About Us	4:44	Xscape	
Thriller Thriller • 1982	4:28	Man in the Mirror Bad • 1987	5:19	Earth Song	6:47	Dangerous Dangerous • 1991	
Billie Jean Thriller • 1982	4:50	You Are Not Alone	5:45	Bad Bad • 1987	4:07	You Rock My World	
Smooth Criminal Bad • 1987	4:17	Heal the World Dangerous • 1991	6:25	Black or White Dangerous • 1991	4:16	Don't Stop 'til You Get Enough Off the Wall • 1979	

News for songs by michael jackson

- Michael Jackson hologram debuts at Billboard Awards ...**
The Globe and Mail - 1 day ago
The late King of Pop Michael Jackson made an appearance as a hologram, performing his posthumously released song 'Slave to the Rhythm' ...
- Billboard Music Awards 2014: Michael Jackson hologram ...**
Telegraph.co.uk - 21 hours ago
- Michael Jackson Billboard Music Award hologram is no ...**
Kansas City Star - 1 day ago

Michael Jackson
1,022,597 followers on Google+

Michael Joseph Jackson was an American singer-songwriter, actor, and businessman. Called the King of Pop, his contributions to music, dance, and fashion, along with his publicized personal life, made ... Wikipedia

Born: August 29, 1958, Gary, Indiana, United States
Died: June 25, 2009, Holmby Hills, Los Angeles, California, United States
Children: Paris-Michael Katherine Jackson, Prince Michael Jackson II, Michael Joseph Jackson, Jr.

Figure 4.6: complex carousel in GKG.

4.2 METHODOLOGY

In this section, we outline the methodology and processes we adopt to investigate the prevalence of list entity results in the semantic search engines.

The evaluation of any information retrieval system requires a standardized approach and control settings which will determine the integrity and validity of the evaluation. There can be many approaches to evaluate information retrieval systems. The method to use in the evaluation has always been challenging. Consequently, the choice of the approach depends on the intent of the evaluation which also determines the type of results obtained [32]. The

Cranfield methodology has been the most widely used approach in evaluating information retrieval systems [74] and has helped in the evolution of other methods. It involves the use of fixed dataset, a set of queries, a set of judges or assessors, and some metrics such as precision and recall [75].

In any evaluation exercise, it is always important to first analyze what to evaluate and the process or procedure to be used in the evaluation. The aim of this section has been to evaluate the prevalence of multiple entity list results for GKG and Satori in answering user's queries. The procedure is to test a set of queries we built by considering 100 classes from Freebase and test them in GKG and Satori. The expected output from the system shall be a list of entities arranged horizontally in a carousel that are possibly ranked from left to right. In the following sections, we explained the methodological approach and settings we used in our evaluation process, which includes the query set, query testing and relevance judgments.

4.2.1 Class Selection

As shown in Figure 3.1 from the previous chapter, the classes in GKG and Satori are unequal. There were 15 classes in Satori that were not found in GKG and 9 classes in GKG that were not in Satori. Conversely, there were 25 classes not found in both engines. Choosing all 100 classes will seem to give an advantage to one of the engines. In view of this, we choose only those classes that are found in both engines to avoid bias. Therefore we used the 51 classes that were in both engines for our evaluation. However, we also tested those classes that are found in either of the search engines. This will allow us to estimate how much of the classes found in either of the search engines have the carousel result (list of entities). For example among the 60 classes found in GKG how many have the carousel result? We therefore tested an overall of 75 classes that are found in either GKG or Satori.

4.2.2 Query Set

We built a query set that will be used to test the presence of list entities in GKG and Satori. These queries are the inputs to the system. A good search query should be able to express the user's intent in a realistic and efficient approach. Queries expressed in a simple and comprehensive way are most likely to produce a better result and are easy to handle by the search engines. It is obvious that how queries are expressed can significantly affect the results in search engines. Therefore, we created queries that are realistic and easy to understand by the search engines. We created a query set from 75 classes that are found in either GKG or Satori out of the 100 classes we randomly generated from Freebase. For each class, we created four queries that would be tested in the semantic search engines. We therefore created 300 queries. The way we formed the queries was to append some adjectives or adverbs to the names of the Freebase classes as a suffix or prefix. The words we appended in the classes are to help fetch some of their instances in GKG and Satori. This is to confirm the availability of list of entities or the instances of the classes in the engines. The kind of words or phrases we append to the classes includes 'top', 'famous', 'list of', 'popular', 'most e.g. most demanded', 'common', 'best' etc. A sample case would be: for the class 'American football game' we constructed queries like: 'Popular American football games', 'top 10 American football games in history', 'famous American football games in history', 'best American football games in history'. For some classes, we added suffix such as dates, location or other phrases. Sample queries we formed using this approach includes: top *musical groups* 2013, famous *musical groups* in America, top *musical groups* in the world. It is obvious that the names of the classes may not always be the best keywords to search for the list of entities in the search engines. For example, for the 'person' class, a search for 'famous person from united states' do not produce any result in neither GKG nor Satori, but searching for 'politicians in United States' have result in GKG. Similarly, for the class 'animal breed', searching for 'list of Animal breeds' or 'famous Animal breeds' has no result in both GKG and Satori, but searching for 'famous dog breeds' or 'cat breeds' have results in both GKG and Satori. How we formulate each query therefore depends on individual class. While formulating the queries, our objective was to come up with queries that can fetch a list of entities from the search engines. It is

possible to therefore refine queries during testing, after gaining some experience from the output of some queries. Table 4.1 shows entity types and their queries.

Table 4.1: Query types and some sample queries

Query Type	Queries
Amusement Ride	<i>Amusement Rides, amusement rides in Las Vegas, Famous Amusement Rides, top amusement rides in the world</i>
House	<i>famous houses, famous houses in Washington dc, famous houses in the world, top 10 most expensive houses in the world</i>
Shopping center	<i>famous shopping centers in London, shopping malls in Los Angeles, Washington dc shopping malls, top 20 shopping malls in USA</i>
Musical Group	<i>top musical groups 2013 , famous musical groups, top musical groups in America, top musical groups in the world</i>
American football team	<i>famous American football teams, top 10 American football teams , top 10 American football teams in united states, NFL teams</i>

The prefix or suffix we appended to the classes sometimes adds some conditions or limitation to the queries which are very important. We therefore identify two kinds of limitations/conditions which are ‘location’ and date. 1) Location: The location added serves as a filter. For example searching for ‘lakes’ has no list of entities from GKG and Satori, but a search for ‘lakes in <country name>’ has result in both search engines. similarly, searching for ‘museums’, ‘revers’ or ‘islands’ has no results in both search engines. But searching for ‘museums in <place name>’ or ‘rivers in <place name>’ or ‘island in <place name>’ produces results in both search engines.2) Date: date may also play a significant role in filtering entities to produce some list of entities in the carousel. For example,

searching for ‘movies’ has no result in both GKG and Satori, but searching for ‘movies 2013’ outputted a list of movies produced in 2013.

4.2.3 Search Engine Settings

In order to be consistent in testing the queries, we used United State Google and Bing engines. This is very important because search results nowadays depend on location. This is why they tend to implement knowledge databases for different countries and languages. *“The complex thing with rolling this out globally is that searching for [chiefs] in the States should return a very different knowledge graph than searching for the same thing in the United Kingdom. As Danny lived blogged: It’s hard to make predictions because the same word can mean different things. Cookies in the US are called biscuits in India, and biscuits mean different things in the US. Oh, how I’ve lived that same issue in Britain. Ask for a biscuit, you aren’t getting a biscuit.”* [76].

4.2.4 Query Testing

All the queries were tested in March, 2014. While testing the queries, there were some challenges we encountered. Sometimes we need to refine the queries in order to produce some results in the search engines. For example, the query ‘Michael Jackson musical albums’ does not produce the carousel result in GKG while ‘Michael Jackson music albums’ displayed Michal Jackson’s musical albums in the carousel. The difference between the two queries is on the word ‘music’ and ‘musical’. This off course is challenging and we therefore in some cases consider many alternatives that we think can produce results in the search engines.

4.2.5 Relevance Judgment

We evaluated the two systems using a two scale metric (1 Or 0). The explanations of the scales are given below.

1 scale: A class shall be evaluated to 1 if at least one of the queries from the class displayed the carousel result that is relevant to the query. For example, both GKG and Satori were evaluated to 1 for displaying a relevant carousel result for the Tourist attraction class query ‘list of tourist attractions’.

0 scales: A class is evaluated to 0 if none of the queries from the class produces a list of entities in the carousel. For example, both GKG and Satori were evaluated to 0 for not having any result for the four queries from the ‘Baseball League’ class. Queries that returned a single entity were also evaluated to this scale. For example, the class ‘Shopping center’ was evaluated to 0 in Satori for outputting ‘Los Angeles’ as the result of the query ‘shopping malls in Los Angeles’, ‘London’ for the query ‘famous shopping centers in London’, ‘Washington, D.C’ for ‘Washington dc shopping malls’ and having no result for the query ‘top 20 shopping malls in USA’.

4.3 RESULTS

The results of our evaluation are shown in Figure 4.7 and Figure 4.8 for GKG and Satori respectively.

4.3.1 GKG Result

From Figure 4.7 it shows that GKG had implemented the carousel result for 10 out of the 51 classes we tested. This means that about 20% of the classes we tested have the carousel result. This result is for the 51 classes that are found in both GKG and Satori from the previous section. The result decline if we only consider the classes/entity types implemented by GKG, with only 10 out of 60 i.e. about 17% of the entity types found in GKG has the carousel implemented.

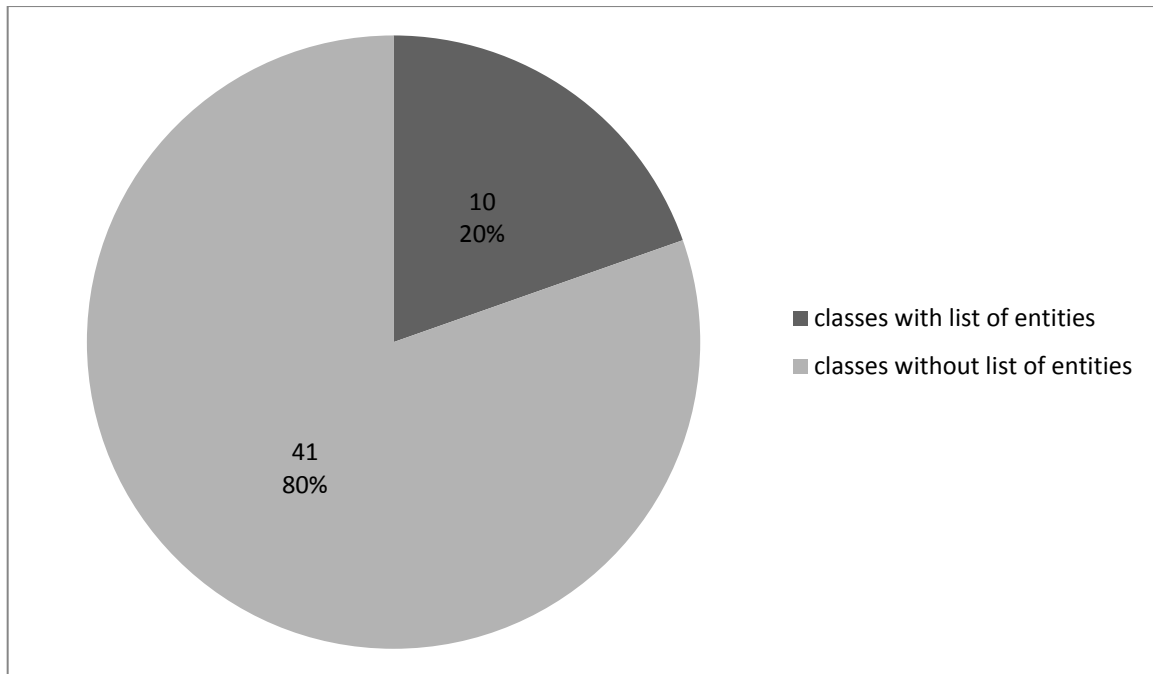


Figure 4.7:GKG result for multiple entity search.

The classes having the carousel includes: *Amusement Rides, Houses, Academic institutions, Musical Groups, Person, Theaters, Tourist attractions, American football teams, Animal breeds and Geological formations.*

The carousel appear in GKG for queries like: *amusement rides in Las Vegas, famous Amusement Rides, top amusement rides in the world, Amusement Rides, famous houses, famous houses in Washington dc, famous universities in UK, top universities in USA, top musical groups 2013, famous musical groups, top musical groups in America, top musical groups in the world, famous person from united states, Famous Politicians, 10 Most Famous American Politicians, famous theatres in the united states, famous theatres in London, list of tourist attractions, tourist attractions, top tourist attractions in the world, famous tourist attractions, NFL teams, famous dog breeds, cat breeds, geological formations in the united states, geographical features of the united states.*

4.3.2 Satori Result

The result from Satori (Figure 4.8) shows that only 7 out of 51 of the classes we tested has the carousel result. The rest of the classes displayed either single entity, map or shows no result. Based on the result we obtained, about 14% of the classes tested have the carousel result. But the result even decline considering the 66 entity types implemented in Satori based on our result from the previous section. Only 7 out of 66 of the classes have the carousel result, which means that only about 11% of Satori classes have the carousel implemented. The classes with the carousel result includes: Houses, Planets, Academic institutions, people, Tourist attractions, American football teams, Animal breed.

The following queries shows the carousel result in Satori: *famous houses: Famous Houses, famous houses in Washington dc, famous houses in the world, Famous Houses, the solar system planets, Solar System Planets, famous universities in UK, top universities in USA, list of tourist attractions, tourist attractions in turkey, famous tourist attractions. NFL teams, famous dog breeds, cat breeds.*

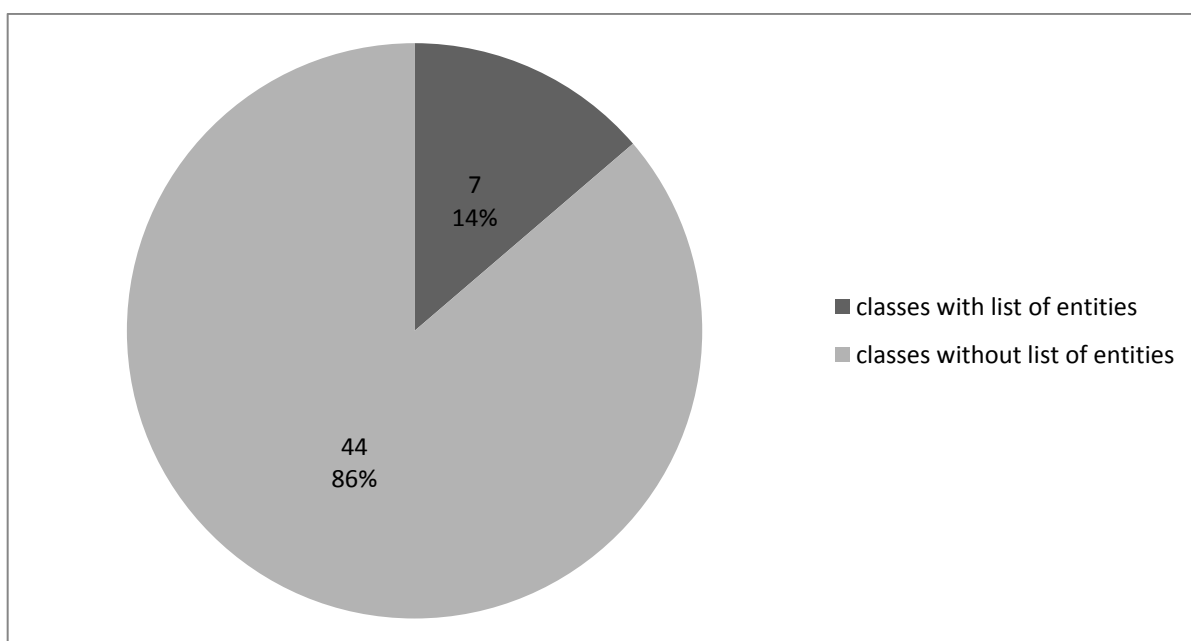


Figure 4.8: Satori result for multiple entity search.

Comparing the two results, it shows that GKG provides more frequent result for searches about list of entities than Satori. While Satori does not have carousel results for the classes: Amusement Rides, Musical Groups, Theaters and Geological formation, GKG does not show the carousel result for the Planets class.

It is also important to note that the result can change in the future as both GKG and satori implement the carousel result for additional classes.

4.4 CONCLUSIONS

Our result on the investigation of list search services by the semantic search engines shows that GKG display list search results more frequent than Satori having shown the carousel result for 20% of the classes tested while Satori shows the carousel results for 14% of the classes tested. The result also shows that GKG has implemented list search services for 17% of its classes while Satori has implemented the list search result for 11% of its classes. The kinds of classes that show results for searches about list of entities include: Amusement Rides, Houses, Academic institutions, Musical Groups, Person, Theaters, Tourist attractions, American football team, Animal breeds, Geological formations and Planets. GKG shows list of entities for all the classes Satori shows except for the “planet” class. While Satori couldn’t shows a list of entities for queries from the Amusement Rides, Musical Groups, Persons, Theaters and Geological formations classes. Other classes that have display list of entities includes: movies, albums, songs, rivers, islands, lakes, mountains, parks, shopping malls, hotels, cities etc.

CHAPTER 5

QUERIES SUPPORTED BY GKG AND SATORI

5.1 INTRODUCTION

Almost every semantic search engine may have different capabilities of the kind of queries it support. An intersection of the kinds of queries supported or handled by a set of semantic search engines would likely produce an “easy” workload (if not empty) [55]. The kind of queries supported by any search engine will determine its capabilities or usefulness to the users. Queries in semantic search may be specified using a formal language or structured query language such as SPARQL, keyword or combination of both [55]. Queries specified using structured query language are more welcome by semantic search engines because of their ability to clearly define the search intent of the user, but they are difficult for ordinary users since a lot of programming is needed to define a query. On the other hand, free text queries are much easier to be used by users but are more problematic for search engines. They require Natural Language Processing (NLP) to understand the intent of users.

5.1.1 Research Questions

In this chapter, we investigate the third research question:

1. What kinds of queries supported by Google Knowledge Graph and Bing’s Satori? What is the extent of their natural language understanding capability? What kinds of natural language queries can they understand and produce correct results? We used constructed

queries about US geography which consist of natural language queries of different complexities.

2. Can search engines understand the most commonly used user queries and return the correct entities? We used the most frequently used 1000 queries of 2008 from Yahoo.

5.2 QUERY INTERFACES OF SEMANTIC SEARCH ENGINES

A good search engine does not just excel by providing relevant results to its users, but should also do a lot to simplify user's interaction. From the search engine vendor's perspective, a choice of how users query the search engine depends on several factors, which may include implementation issues such as usability, portability, simplicity etc. Each method may have its advantages and disadvantages on both the users and the search engine. In essence, the method used by users to interact with the search engines have a great impact on the quality of the result they can obtain, since most users do not know exactly how to express their search needs. In the next few paragraphs, we stated the methods of interactions with the semantic search engines, pointing out the merits and demerits of each method.

5.2.1 Structured query languages

Structured query languages such as SPARQL [20] and Cypher [21] let users to describe the information needs precisely. For example following queries can be formulated easily: "10 largest cities in the united states", "actors or actresses that played in at least 2 movies in 2012". However, these languages require users to know the structure of entities and the relationships in the database. In addition, they require users to formulate their information needs by using a formal query language [22]. Search engines do not want to make their entity types and relationships public. In addition, it is very difficult for general users to use a formal language to describe their information needs. Therefore, current search engines do not prefer structured query languages as their query interfaces for general users.

5.2.2 Form-based

In form-based semantic systems, users are presented with a form consisting of menus, drop-down boxes, check boxes etc. Users specify their queries by filling, checking, or selecting the items in the form menus. In this case, mapping of user's key-words to ontologies in the entity database is avoided since users queries are restricted to selected terms that are in the system's database or record. By displaying the available terms and features in a form, users can get an insight of how the domains are structured, what constitute the domain, and what kind of queries they get from their queries. Form based interface may be ideal for some semantic search engines because of the hierarchical structure or organization of entity types in semantic web data. On the contrary, form-based interface are not flexible. They restrict their users to a limited options or items to select which are guarantee to be available in the system database, thereby disallowing them to perform an exploratory search. Moreover, form-base may not be feasible as the entity database becomes larger and complex, since some form fields may have a limitation on the number of field items or values that are selectable. Moreover, most of the form-based systems have no heterogeneity (searching in multiple ontologies) implemented. Searching across multiple ontologies is particularly important in semantic searching and therefore, its absence in form-based systems will limit their applicability [56].

5.2.3 View based

In view-based systems, users are presented with a graphical or tree like interface for which they can navigate in or out of a set of ontologies to find their information need. The advantage of view-based systems is that the ordering of the ontologies and classification gives users an intuitive format which allows them to grab an understanding of the entity domains. In addition, view-based systems also give users the ability to construct complex queries without knowing the domains or structure query languages since users perform queries by clicking or navigating through the ontologies. A user start querying by selecting a domain class or entity type, the properties or subclasses associated with the class are being listed and users may expand their queries by selecting those properties or subclasses.

This continues until users are able to reach their target. Each navigation or expansion by user generates a different query and the deeper a user navigate into the ontologies, the more complex the query will be. The main disadvantage of this kind of systems or query interfaces is the time consumed before users reach their target and it is not flexible. Also, Heterogeneity (combining multiple ontologies from different domains in query) is another problem encountered by view-based systems.

5.2.4 Natural Language Queries

Traditionally search engines have been using natural language queries to serve users. Compared to the formal query languages, these queries may define inexact entities and concepts, and easier to use by users [23]. In addition, there are well known algorithms to match the query strings to documents. Without trying to understand the meaning of the words, these algorithms measure the similarity of query strings to document contents. Nonetheless, querying entity databases requires semantic parsing of query strings. When executing the query “capital of Germany”, the word “Germany” needs to be identified as a country and the word “capital” needs to be identified as the administrative center of it. Similarly, when executing the query “books by Obama”, the word “Obama” needs to be identified as the author Barack Obama and the word “books” need to be identified as published works of him.

Semantically parsing natural language sentences are difficult, since natural language sentences involve variability and ambiguity [28]. The same statements can be made in many ways and one sentence may mean multiple things. There have been various studies to query semantic web datasets with natural language interfaces. Many of these systems either limit the input language to a subset or work on domain specific datasets to reduce the variability and the ambiguity in the input queries [29], [30]. It is really difficult to develop unlimited natural language interfaces to domain independent datasets with high precision and recall.

5.3 RESULTS RETURNED BY SEMANTIC SEARCH ENGINES

Traditional search engines return a ranked list of documents related to user's queries. However, there is no standard way of presenting the results of semantic web search engines. For semantic search engines, an entity or collections of entities satisfying the user's query and/or their relationships to other entities are expected to be returned as the result for users' queries. In particular, the result returned by semantic search engines is dependent on the type of query. In the case of GKG and Satori, results presented from their entity database are of three categories:

5.3.1 Single Entity

For searches about a particular entity, the entity is. It usually has a picture and some properties that may be of an interest to users. In addition, a set of related entities are suggested to the user which define how the entities are related with the entity being searched. Figure 5.1 shows the result of searching 'president Obama' in Google. The entity for Barack Obama is shown on the right hand side of the SERP.

The image shows a Google search result for the query "president obama". The search bar at the top contains the text "president obama" and a search icon. Below the search bar, there are navigation tabs for "Web", "Images", "News", "Videos", "Books", "More", and "Search tools". The search results show "About 229,000,000 results (0.38 seconds)".

On the left side, there are several search results:

- News for president obama**: A list of news articles, including "Samsung, selfies and the branding of Barack Obama" from the Washington Post (blog) by Jaime Fuller, 4 hours ago, and "U.S. President Barack Obama poses with player David Ortiz for a 'selfie' as he welcomes the 2013 World Series Champion Boston Red Sox to ...".
- TRANSCRIPT: President Obama's remarks on Fort Hood shooting** from the Washington Post by Mark Berman, 8 hours ago.
- Column: Waiting for President Obama at Zingerman's Deli** from The Ann Arbor News, 5 hours ago.
- More news for president obama**
- Barack Obama - Wikipedia, the free encyclopedia**: A link to the Wikipedia page for Barack Obama.
- President Barack Obama | The White House**: A link to the White House website for Barack Obama.
- Barack Obama**: A link to the Barack Obama website.
- Barack Obama (BarackObama) on Twitter**: A link to Barack Obama's Twitter profile.

On the right side, there is a knowledge panel for **Barack Obama**. It includes a photo of Barack Obama, a "Follow" button, and the following information:

- Barack Hussein Obama II** is the 44th and current President of the United States, and the first African American to hold the office. [Wikipedia](#)
- Born:** August 4, 1961 (age 52), Honolulu, Hawaii, United States
- Spouse:** Michelle Obama (m. 1992)
- Office:** President of the United States since 2009
- Parents:** Ann Dunham, Barack Obama Sr.
- Siblings:** Malik Abong'o Obama, Maya Soetoro-Ng, More
- Children:** Natasha Obama, Malia Ann Obama

Below the knowledge panel, there are "Recent posts" and "People also search for" sections. The "People also search for" section includes images and names of Michelle Obama, George W. Bush, Vladimir Putin, Hillary Rodham Clinton, and Bill Clinton. A tooltip at the bottom of the knowledge panel states: "Barack Obama and Michelle Obama have been married since 1992."

Figure 5.1: Single entity result.

The entities suggested to the user in the ‘people also search for’ section may have some ranking mechanism. The entities suggested in this section are not fixed and may be based on previous and current events happening between entities in the real sense.

5.3.2 Set of Entities

For a search about set of entities, which satisfies a given criteria, a panel of entities is presented to the user at the top of the SERP. In this case, the entities are listed without their properties. However, a click on each entity on the panel displays the entity together with its properties on the right hand side of the SERP. Figure 5.2 shows a search for ‘Texas cities’ in Bing.

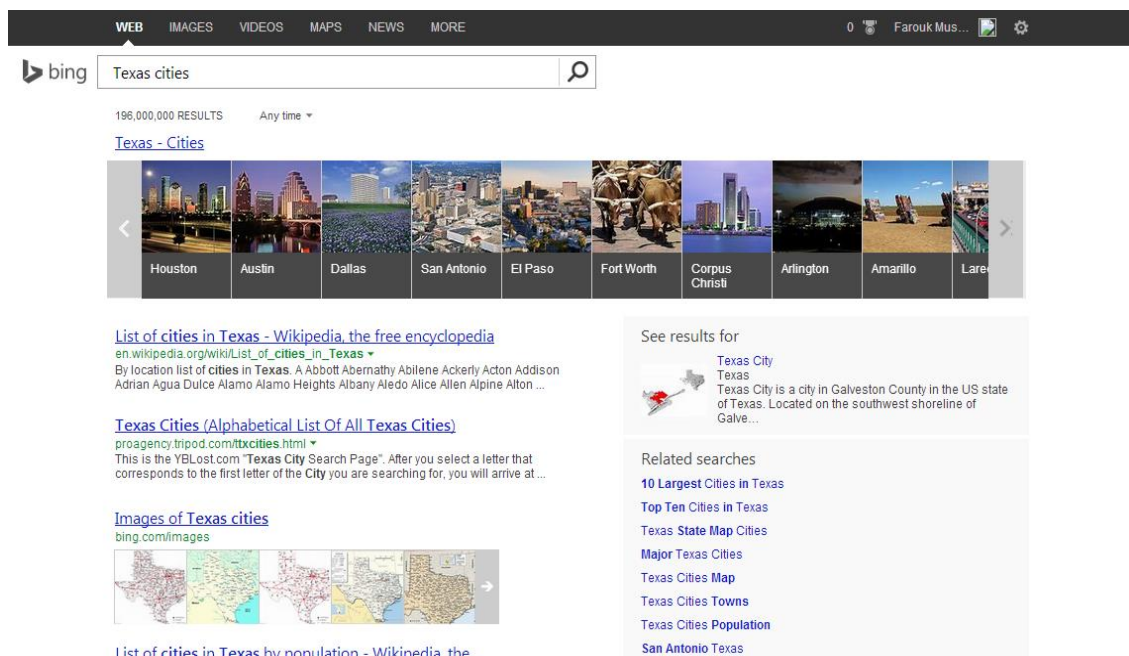


Figure 5.2: multiple entity result in Satori.

5.3.3 Attribute of Entity

For searches about an attribute of an entity, the entity itself is displayed at the right hand side of the SERP, and the value of the attribute is usually displayed at the top of the SERP. Figure 5.3 shows the Mount Everest entity on the right side of SERP and the height of Everest at the top of the result page for the query ‘‘height of mount Everest’’.

Google height of mountain everest

Web Images News Shopping Maps More Search tools

About 1,150,000 results (0.36 seconds)

29,029' (8,848 m)
Mount Everest, Elevation

Mount Everest

Mount Everest is the Earth's highest mountain. It is located in the Mahalangur section of the Himalayas. Its peak is 8,848 metres above sea level and is the 5th furthest point from the centre of the Earth. Wikipedia

Elevation: 29,029' (8,848 m)
First ascent: May 29, 1953
Prominence: 29,029' (8,848 m)
First ascenders: Tenzing Norgay, Edmund Hillary
Mountain range: Himalaya, Mahalangur Himal

Mount Everest - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mount_Everest * Wikipedia *
Mt. Everest from Gokyo Ri November 5, 2012 Cropped.jpg ... Trigonometric Survey of India established the first published height of Everest, then known as Peak ...
List of people who died ... - List of mountains - Tenzing Norgay - Edmund Hillary

7 Things You Should Know About Mount Everest — Hist...
www.history.com/.../7-things-you-should-know-abo... * The History Channel *
May 29, 2013 - Explore some surprising facts about Mount Everest 60 years after it was ... and by 1856 they had calculated its height as 28,002 feet above sea ...

Height of Mount Everest (Everest, Mount) -- Encyclope...
www.britannica.com/.../Height-of-Mount-Ever... * Encyclopaedia Britannica *
The height of Mount Everest, according to the most recent and reliable data, is 29,035 feet (8850 metres). In 1999 an American survey, sponsored by the (U.S.) ...

Facts About Mt. Everest - Scholastic
teacher.scholastic.com/activities/hillary/.../evefacts.htm * Scholastic Press *

Figure 5.3: Attribute of an entity result in GKG.

As we mentioned above, the display of results by semantic search engines doesn't have any definite format or standard. It is obvious that the way GKG and Satori display their result may change in the future, depending on what may be convenient for their users.

5.4 OBJECTIVE OF THE CHAPTER

The purpose of this chapter is to investigate the query types supported by GKG and Satori, and the capabilities of the semantic search engines by using both constructed and real user's queries. Our results may be helpful in two ways: (1) to the search engine users, the result will help them capitalize on their searches by knowing the kinds of queries supported by the search engines and how strong are the search engines in recognizing their inputs. (2) To developers of semantic search engines and to the researchers on this field, it shows the capabilities of current semantic web search engines.

5.5 METHODOLOGY

Even with the growth of the semantic web and the evolution of semantic search engines, there has been little effort or research towards a standardized methodology for evaluating semantic search engines. The first systematic evaluation of semantic search engines which uses information retrieval metric was in 2010 [55], but it suffers many criticisms due to its query simplicity. Recently, Jeffrey Pound et al. in their paper “Ad-hoc Object Retrieval in the Web of Data” [55] propose a mechanism for evaluating object retrieval systems that modifies the traditional Ad-hoc Document Retrieval (ADR) task. The Ad-hoc Object Retrieval (AOR) they proposed is as follows:

- **INPUT:** a user query (a keyword query, without structure) q which has query type t and query intent z , and a data graph G .
- **OUTPUT:** a ranked list of resource identifiers $o = (o_1, o_2, \dots, o_k)$ such that each o_i occurs in G .
- **EVALUATION:** each object (or resource) o_i is labeled with a score (independently of the rest) by a judge with access to all the information contained in or linked to by o_i , with respect to the query q , query type t , and the query intent z .

The steps above can be seen as generalized steps for evaluating an entity retrieval system, but each evaluation exercise may have a different set of input, output and evaluation metrics. Each evaluation exercise have its own type of input which consist of the queries that will be used for testing the system, the output depends on the system to be evaluated and each system have its own type of output and capabilities. The quality of the evaluation will depend on the evaluators or judges.

We basically adapt to the steps above with some minor modifications. We use Geo-queries from [54] as inputs which are natural language queries and also real user queries from Yahoo 1000 most frequent web search queries issued to Yahoo! Search over a three month period in 2008, (English language, version 1.0).

Semantic engines are design to improve search accuracy and understand user's intent. They try to limit the number of irrelevant results returned to the user. Most semantic engines engaged in many activities to improve their search result which includes: the context of search, location, current trend and news, Intend of the search, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries [62]. Current semantic web search engines usually return one entity for majority of queries. In view of this, it will be of utmost important to return the correct entity. However, this is still a big challenge for the semantic search engines.

5.5.1 Search Engine Settings

In order to be consistent in testing the queries in both GKG and satori, we set both engines to their US interfaces. We use www.google.com for GKG and www.bing.com. In the case of Satori, we also change the location to "United State-English" from the settings in Bing's home page. Users can change their location by clicking the settings icon by the top right corner of the search engine's home page. Inside the settings, the location can be changed by clicking the country/region link and then selecting United State-English. This is very important because the queries we used are about US Geography and users may not get result without the settings especially in the case of Bing. Also GKG or Satori may not be available in some countries, as they both reported to have started from some few countries. Moreover, we predict that the location used in searching can also affect the quality of the result obtained in some semantic search engines.

5.5.2 Query Sets

Two classes or sets of queries were used in the evaluation, the first are constructed queries (Geo-queries) and the second are real query log. Our objective of using different sets of queries is to test the capabilities of the semantic search engines thoroughly since the query sets has different complexity. The queries also consists of the different classes of query types such as the entity queries, multiple entities queries, attribute queries etc. which are good for testing semantic search engines.

We aimed at achieving different objectives by using constructed and real user queries. The constructed queries are not only aimed at testing the types of queries supported by the search engines, but also the natural language understanding capacity of the search engines. They consist of good natural language queries of various complexities. The major side effect of using the constructed queries is that they are domain specific i.e. they only cover one domain (Geography). It is obvious that a search engine may perform better in some domains than the others, as the quality of results from search engines depends on the amount of records in the particular domain. For example, the quality of result about restaurant search will depend on how much information/records (entities and their relationship) GKG or Satori has accumulated about restaurants. In view of this, we do not use the constructed queries to major the performance of the search engines but how they understand natural language queries.

In the real query log, the queries consist of queries from different domains and the natural way users query search engines. Therefore, these queries are good for testing the overall capabilities of the semantic search engines in responding to users' queries. The result obtained from this test can therefore be a generalized performance measurement for the search engines.

5.5.2.1 Geo-Queries

These queries consist of about 880 constructed natural language queries about US geography that was gathered from different source. It includes queries about states, state capitals, cities, population, rivers, lakes etc. 250 of these queries were gathered from undergraduate German language class. No instructions were given to limit the complexity of the queries. The remaining 630 queries were gathered from undergraduate AI class and from users of a web interface to a CHILL [54] prototype trained on the initial 250 data set [57]. The AI students tended to asks more complex and diverse queries. The initial 250 queries from the German class students were translated to English and together with the 630 queries from the AI students form 880 queries. These queries are good for entity retrieval evaluation and consist of various degrees of query complexities including simple,

moderate and complex queries. All the queries consist of one or more semantic resources including entities, types, attributes and some statistical results. The geo-queries has been used in variety of researches including semantic parsing [57], semantic search evaluations study [58], Semantic Web Natural Language Interface (NLIs) studies [59], [60], [61] and many other studies relating to semantic searching and NL.

5.5.2.2 Real Query log

This query set consists of 1000 most frequent searched queries issued to Yahoo! within three month period in 2008. The dataset consist of nine files each of different language and consisting of queries based on the most frequent queries issued to Yahoo! Search. The queries are provided as part of Yahoo! Webscope program available for usage to non-commercial academic researchers under the terms and conditions of a signed Yahoo! Data Sharing Agreement [79]. Majority of the queries are searches about single entities including people, companies, places, organization, social network site, movies, games, software products, etc. Each query contains a word or few words. The dataset consist of queries that search about entities users are most interesting in the real world. These queries are good for entity search evaluations. The dataset has been used in quit a number of researches which includes entity search evaluations [43] and automatic semantic content creation [80].

5.5.3 Query Types

The categorization of queries into types is a major step to evaluate semantic search engines. Each category has its own kind of results which tell what is/isn't a valid result for a query. This classification is not necessary when evaluating keyword or traditional search engines because the result returned by these engines are always a list of documents. The classification of queries into types is independent of the search engines but depends on what it's expected to return or the intent of the query. Therefore, it is possible to have several query types, depending on each query and what is expected to be returned. We

manually examined each query and determined their category. A research to classify queries for entity search was conducted by [55]. They manually examined a real web search engine query log and they were able to categorize the queries into five distinguishing types: (1) Entity Queries which search for a specific entity, (2) Type Queries which search for multiple instances of a particular type or class, (3) Attribute Queries which search for the attribute value of an entity, (4) Relation queries which search for the relationships between two entities, and (5) other keyword queries which are keyword queries that do not fall under any of the above categories. Another study [19] confirms that users may perform three of the following activities in relation to entity search: (1) search for a specific entity, (2) Search for a set of entities (3) Search for an attribute of an entity. The study also gives some statistics about each category, with search for single entity queries having a dominant percentage of about 60% - 70% and 10-20 % for sets of entities.

In our study, we determined five types of queries. We manually annotated each query to one of these query types. The details of each query type are given with example queries:

- **Single entity:** The intention of this kind of queries is to find an entity E.g. “What is the capital of Texas”.
- **Multiple entities:** the intention of this kind of queries is to return a set of entities that satisfy the query. E.g. “Give me all the states of USA”.
- **An attribute of an entity:** The intention of this kind of queries is to find the attribute of an entity E.g. “What is the population density of South Dakota?”
- **Attributes of multiple entities:** Some queries may ask for the attributes of multiple entities such as “What are the populations of the major cities of Texas?”
- **Statistical queries:** These are queries that return a statistical operation either on group of entities or their attributes. E.g. “What is the total population of the states that border Texas?”

The number of queries in each query type for the geo-queries queries is shown in Table 5.1.

Table 5.1: The number of queries in each query type for the Geo-queries queries

<i>Query Type</i>	<i>No. of Queries</i>
Single Entity	375
Multiple Entities	212
Attribute of an Entity	184
Attribute of Multiple Entities	17
Statistical Queries	89

The categorization of queries into types is very important, since each query would return an answer depending on its type. Therefore, queries that do not return answers based on their query types were penalized.

It is not easy sometimes to decide the query type of some queries. For example, the query ‘How many rivers are in New York’ should it be consider as multiple entities type or statistical query? According to the context and grammar of the query, the query should be of type ‘statistical query’, since it return a number (frequency of rivers in New York). But most semantic search engines do not perform statistical operation on their data and may treat the query as a multiple entity type by just returning the list of rivers in New York. Making this kind of decisions is quiet challenging.

5.5.4 Evaluation Scale

We evaluated each result from GKG and Satori using two scales (0 Or 1). The explanations of the scales are given blow.

1 Correct: For a result to be evaluated on this scale, it must satisfy the intent and the type of the query. For example the query ‘how many people live in Washington?’ is of type attribute of an entity, and it’s intended to output the population of Washington. Both GKG and Satori outputted Washington and the figure D.C: 632,323 (2012) in their regular SERP, which we strongly believed to come from their entity database.

0 incorrect: the 0 scale here have several intuitions, which includes incorrect or irrelevant results, type mismatch or no result. For incorrect or irrelevant results, the output of the query does not answer the intent of the query. For example, Satori outputted ‘Mississippi River’ for the query 'What are the major cities in states through which the Mississippi runs', the intent of this query is to retrieve the cities in the states through which Mississippi runs and not Mississippi river. The primary entities are the cities and Mississippi is only a secondary entity in the query. Similarly, GKG outputted ‘New York’ for the query ‘what are the major cities in New York?’. If we are to only measure the relevancy of the result to the query without paying attention to the context and intent of the query, we would have marked the output as relevant result, but the output do not match the intent of the query, which are the cities in New York. For type mismatch, the output of the query did not match the type of the query. For example, the query 'What is the biggest river in Illinois?' is a single entity type, but GKG outputted a panel or list of rivers in Illinois which are multiple entities. Other kinds of query mismatch also occurred, but the most common ones are single-multiple entities type mismatches and multiple-single entity type mismatches. Moreover, some queries may have no result in GKG and Satori, for example, as of February 2014, the query 'What is the biggest city in USA?' has no result in both GKG and Satori. Table 5.2 shows the interpretation of the scales with respect to each query type.

Table 5.2: Interpretation of scales

Query Type	1: meaning	0: meaning
Single Entity	If the single entity asked was correctly returned by the search engine.	If: 1) no entity was returned 2) incorrect entity was returned 3) the result returned does not match with the query type.
Multiple Entity	If a panel or list of entities were returned that satisfies the query.	If: 1) no entities were returned 2) incorrect entities were returned 3) the result returned does not match with the query type.
Attribute of an Entity	If the attribute of the entity asked was correctly returned by the search engine.	If: 1) nothing was returned 2) incorrect attribute was returned 3) the result returned does not match with the query type.
Attribute of multiple Entities	If the attributes of the entities asked were correctly outputted by the search engine.	If: 1) no attributes were returned 2) the incorrect attributes were returned 3) the result returned does not match with the query type.
Statistical query	If the search engine correctly returned a calculated figure of what was asked.	If: 1) nothing was returned 2) the incorrect figure was returned 3) the result returned does not match with the query type.

5.5.5 Categorization of Geo-Queries Based On Their Complexities

We categorize the Geo-queries into four categories with respect to their complexities:

- **Simple:** the query should have a simple grammatical structure with one verb and unambiguous intent. The query should target a single entity, a single attribute or a set of

entities belonging to a class and satisfying only one condition. A few examples would be: “What is the capital of Utah?”, “What state is Austin in?”, “How big is Alaska?”, “How long is the Colorado river?”, “What are all the rivers in Texas?”, “What are the cities in California?”, etc.

- **Moderate:** the query should have one or two verbs. In addition, it should have one keyword that requires conditional or selective processing. The keyword may specify an ambiguous intent (major cities), or requires determination of max/min values (largest city, longest river) or requires calculation of a conditional statement (states bordering Texas). A few example queries would be: “What are major rivers in Texas?”, “What is the biggest city in Arizona?”, “What is the biggest state?”, “What is the capital of the smallest state?”, “What state has the most cities?”, “States bordering Iowa?”, etc.
- **Complex queries:** The query shall require two conditional or selective processing. It can have one or two verbs. It may have compound or nested grammatical structure. A few examples: “How high is the highest point in the largest state?”, “What are the highest points of states surrounding Mississippi?”, “Which states does the longest river cross?”,
- **More Complex Queries:** The query shall require more than two conditional or selective processing. A few examples: “What states border states that border states that border Florida?”, “What are the largest cities in the states that border the largest state?”,

5.5.6 Categorization of Yahoo Queries

While testing the yahoo queries, we followed the same procedure as for the Geo-queries. But this time, we do not consider the complexity of the queries. The Yahoo queries consist of one or few words like “Facebook”, “orange”, “internet explorer”, “yahoo games”, “the dark knight”, “windows media player”, “www weather com” etc. Almost all of these queries are single entity type. We shuffle the queries and then select the first 200

queries for our evaluation. Among these 200 queries there are some queries that do not match to the type of queries we needed in evaluate the search engines. For this reason, we filter the queries to only entity search queries that are unambiguous. The queries we filter falls under the following categories which are explained blow.

- **Informational Queries:** These are queries that users ask to find information about things that are not necessarily entities. They do not directly search for a particular entity or refer to any entity. They tend to seek for enquiry about things and usually consist of a noun and an adjective. Examples of these kinds of queries includes: ‘cheap airline tickets’, ‘love quotes’, ‘car insurance’, ‘myspace backgrounds’, ‘myspace codes’, ‘music videos’, ‘hairstyles’, ‘news’ etc.
- **Pornographic Queries:** Among queries that were filtered include queries that searched about pornographic items or websites. A few example of these queries includes: ‘porn hub’, ‘badjojo’, ‘adult friend finder’, ‘youporn com’, ‘brazzers com’ etc.
- **Ambiguous queries:** These are queries that do not have a clear intent. They are queries that include more than one entity. Example of these kind of queries are: ‘yahoo map quest driving directions’, ‘google mapquest driving directions’, and ‘google mapquest driving directions google maps’.
- **Single Alphabet:** some queries may be made of a single alphabet. Single alphabets are not usually used as names of entities but may be an abbreviation of an entity such as ‘H’ for Hydrogen. However, searching for a single alphabet may be confusing and unclear for the search engine. Moreover, a single alphabet may not convey any interesting issues. In the case of our evaluation, we omitted queries of single alphabets.

Table 5.3 shows the number of the query categories we filter during the evaluation. A total of 51 queries were filtered or disqualified. The remaining 149 were tested in GKG and Satori.

Table 5.3: Number of queries filtered from each query categories.

Query Category	Number of Queries
Informational Queries	22
Pornographic Queries	22
Ambiguous queries	3
Single Alphabet	4

5.6 EXPERIMENTAL RESULTS:

5.6.1 Geo-queries Result

All the Geo-queries were tested in GKG and Satori, and the results from both engines were obtained and recorded. The test was conducted in February, 2014. We have observed that the results may change over time. Both GKG and Satori have been expanding their entity databases. Some searches that do not produce any result currently may produce one in the future. Similarly, some searches that currently produce results may not produce any in the future or may output a different result. Figure 5.4 shows a cross-section of the Geo-queries and their results in GKG and Satori.

	A	B	C	D	E	F	G
1	Query	Query Complexity	Query Type	GKG Result	Answer In GKG	Satori result	Answer In Satori
2	Can you tell me the capital of texas?	1	single entity	0	suggested: Texas State Capitol	0	No Result, But show Austin for "the capi
3	Could you tell me what is the highest point in the state of oregon?	3	single entity	0	No Result	0	No Result
4	Count the states which have elevations lower than what alabama has?	3	statistical	0	No Result	0	No Result
5	Give me all the states of usa?	1	multiple entiti	0	No Result	0	No Result
6	Give me the cities in texas?	1	multiple entiti	0	Texas	0	No result, but shows a panel for a search
7	Give me the cities in usa?	1	multiple entiti	0	No Result	0	No result, but show a panel for a search
8	Give me the cities in virginia?	1	multiple entiti	0	Virginia:	0	No result, but shows a panel for a search
9	Give me the cities which are in texas?	1	multiple entiti	0	No result	0	No result, but shows a panel for a search
10	Give me the lakes in california?	1	multiple entiti	0	No result	0	No result, but shows a panel of entities
11	Give me the largest state?	2	single entity	0	No result	0	No result, shows Alaska for a search of "
12	Give me the longest river that passes through the us?	3	single entity	0	No result	0	No result
13	Give me the number of rivers in california?	2	statistical	0	No result	0	No result
14	Give me the states that border utah?	2	multiple entiti	0	No result	0	No result
15	How big is alaska?	1	attribute of an	1	Alaska: 663,300 sq miles (1,718 m	1	Alaska: 663,267 sq miles (1,717,854 sq km
16	How big is massachusetts?	1	attribute of an	1	Massachusetts: 10,554 sq mile	1	Massachusetts: 10,554 sq miles (27,3
17	How big is new mexico?	1	attribute of an	1	New Mexico: 121,697 sq miles (3	1	New Mexico: 121,697 sq miles (315,194 s
18	How big is north dakota?	1	attribute of an	1	North Dakota: 70,762 sq miles (18	1	North Dakota: 70,762 sq miles (183,272 s
19	How big is texas?	1	attribute of an	1	Texas: 268,820 sq miles (696,241 l	1	Texas: 268,820 sq miles (696,241 sq km)
20	How big is the city of new york?	1	attribute of an	1	New York City: 468 sq miles (1,21	1	New York: 469 sq miles (1,213 sq km)
21	How high are the highest points of all the states?	3	attributes of m	0	No result	0	No result
22	How high is guadalupe peak?	1	attribute of an	1	Guadalupe Peak: 8,750' (2,667 m)	1	Guadalupe Peak: 8,751 feet (2,667 me
23	How high is mount mckinley?	1	attribute of an	1	Mount McKinley: 20,322' (6,194 m)	1	Mount McKinley: 20,237 feet (6,168 metr
24	How high is the highest point in america?	2	attribute of an	0	No result	0	No result

Figure 5.4: Geo-queries and the results in GKG and Satori.

The results of the evaluation carried out using the Geo-queries are shown in Figure 5.5 for GKG and Figure 5.6 for Satori.

5.6.1.1 GKG Result

From Figure 5.5, it shows that only the 'attribute of an entity' query type (third bar) has correct results that outnumber the incorrect results, even though the number of correct results in the single entity typed are also expected to be higher than the incorrect results because a vast majority of entity search are single entity typed. The main reason for the high number of incorrect results in the first bar was the complexity of those queries compared to the queries in other query types. Most queries of the type 'attribute of an entity' were about the population of a state, size of a state, height of a mountain and length of a river, and they were expressed in their most natural format. For example 'How big is Alaska?', 'How large is Alaska?', 'How long is the Ohio river?', 'How many people live in Austin?', 'How tall is mount McKinley?', 'What is the population of California?' etc. It is obvious that the way queries are specified has a great impact on the results. Also, the same

query can be given in many ways which may have different complexities. Moreover, some queries may be easy for the system to handle, while others expressed in an unusual way may be difficult for the system. For example the query ‘can you tell me the capital of Texas?’ has no result in both GKG and Satori, but the query ‘the capital of Texas’ gives Austin in Satori. Both queries have the same intent, but the initial is more complex and thus more difficult for the system.

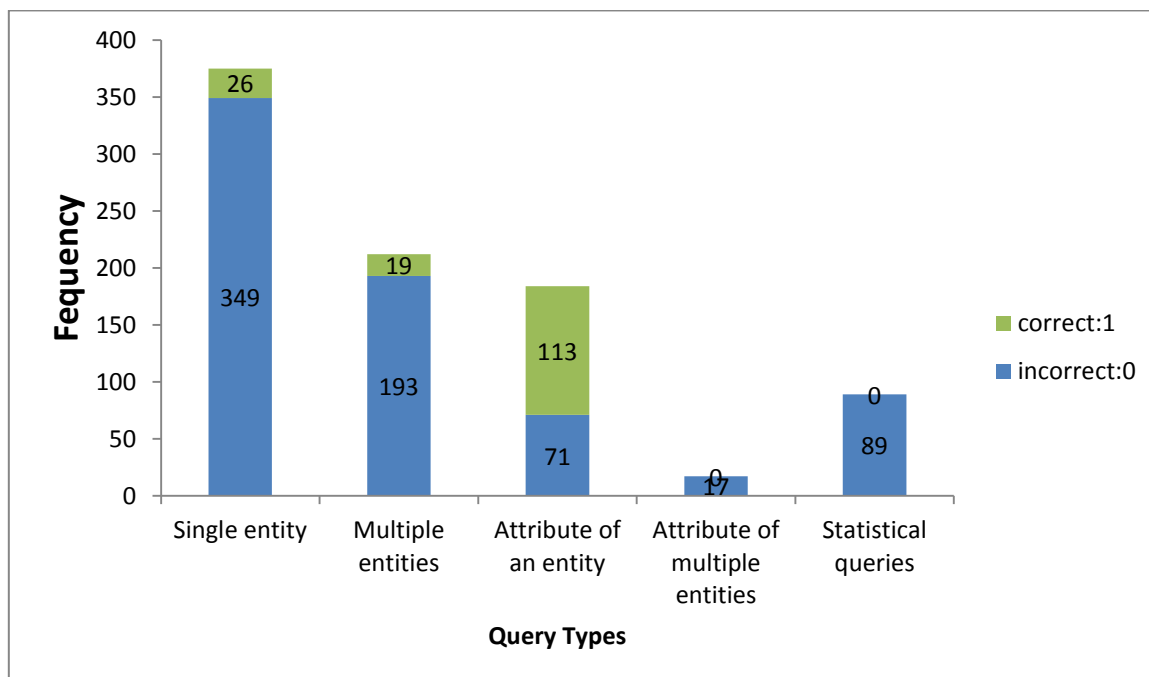


Figure 5.5: GKG Correct and incorrect result for each query type from Geo-queries.

In the single entity type queries, only 6.93% of these queries were correctly answered by GKG. It is obvious that GKG could not get the intuitive meaning of most of the sentences/queries after parsing. The only queries it was able to get correctly are queries about the capitals of US states e.g. 'what is the capital of California?' GKG could not even get some queries about some state capitals expressed in a different way like ‘Can you tell me the capital of Texas?’ and 'What is the capital city of the largest state in the US?'. In total, 87.2% of queries belonging to this query type have no result in GKG.

For the multiple query type, only 8.96% of the queries were correctly answered by GKG, while 78.3% of the queries show no result in GKG. These queries are of utmost importance, they fetch entities from the entity database that usually certify some criteria.

The result is even worse for the statistical and attribute of multiple entities typed. The statistical queries may sometimes require some arithmetic on some of the semantic resources, for example ‘Count the states which have elevations lower than what Alabama has?’ or ‘Give me the number of rivers in California?’. In the attribute of multiple entities type queries, the attribute of multiple entities are required to be displayed by the search engines. The search engine is required to retrieve the attributes of each entity, and then output a combination of these attributes. For example, in the query ‘what are the population densities of each US state?’ the population densities of all US States are required to be return, the population density being an attribute of a state. From our result, it is obvious that GKG is not implemented to return the results of such kinds of queries (attribute of multiple entity and statistical queries).

There are a lot of habitability problem [81] (*The mismatch between the users’ expectation and the capability of the natural language system*) [23] especially among multiple entity typed queries. For example, both GKG and Satori return ‘Alabama’ for the query ‘what are the major cities in Alabama?’. In this case, the user’s expectations are cities of Alabama and not the state of Alabama. This type of mismatch or problem is common among the current NLP tools [23]. Obviously, the cities of Alabama are entities in both GKG and Satori databases but because both systems failed to understand the intent of the query, Alabama was returned instead. This kind of problem is what Google and Bing must have been working on to improve their search results.

For the Geo-queries, GKG has 18.13% accurate (labeled 1) result which means that 81.87% of the queries were inaccurate (labeled 0), 74.91% of the queries have no result from GKG and 6.96% of the Geo-queries have results that either mismatch the query type or incorrect results.

5.6.1.2 Satori Result

The Geo-query result for Satori (Figure 5.6) has a lot of similarity with that of GKG

result. Most of the result in the attribute of an entity type queries for both search engines were similar, with Satori lagging behind GKG only in 3 queries; while GKG get the 3 queries right, Satori outputted the entity ‘Mississippi river’ for the query ‘How long is the Mississippi?’ instead of the length attribute of the entity, which is a query type mismatch. Also, Satori outputted Georgia for the query ‘how many people live in the capital of Georgia?’ which suffers from two problems, one is the habitability problem (the expectation of the user is on Atlanta, the capital of Georgia and not Georgia), secondly, there is query type mismatch, the query is seeking for the attribute of an entity (population of Georgia’s capital) and not an entity. The third query result from Satori was penalized for query type mismatch; Satori returned the entity Texas for the query ‘What is the area of the Texas state?’ instead of the size attribute of Texas.

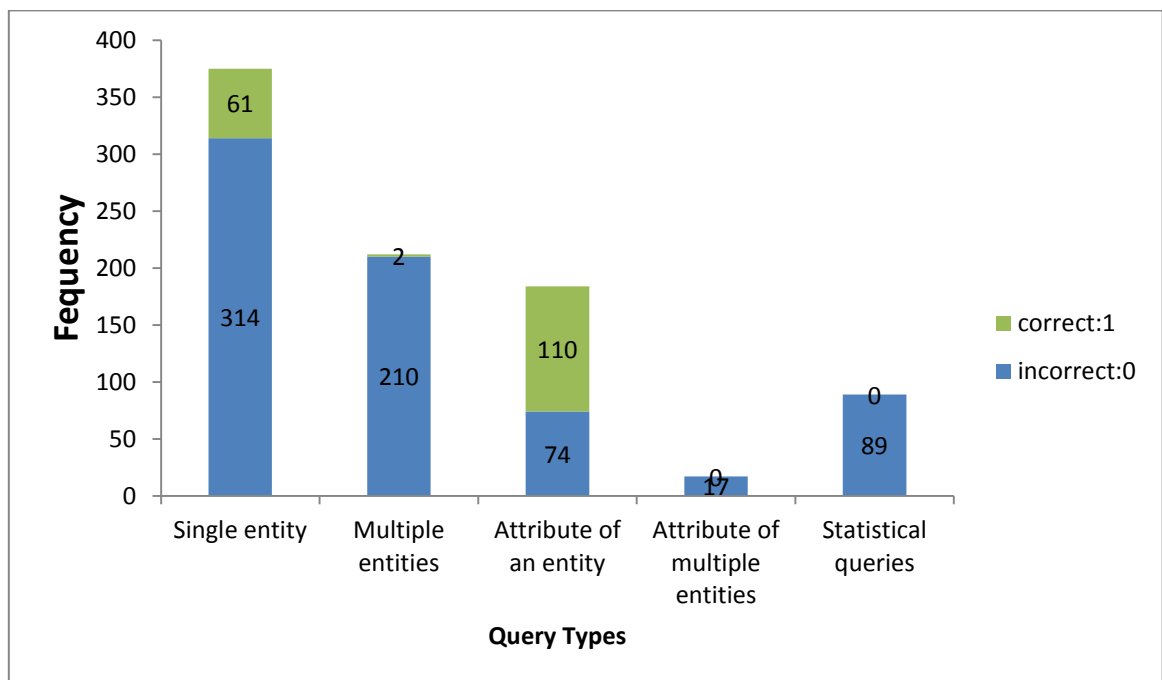


Figure 5.6: Satori Correct and incorrect result for each query type from Geo-queries.

Both Satori and GKG have similar results for the attribute of multiple entity and statistical query type (fourth and fifth bar), with none of the results from either side satisfying the condition for correctness.

The major disparity in the GKG and Satori result is on the single and multiple entity types. While GKG recorded a larger accuracy of results in multiple query type than Satori with 19 accurate results, and Satori having only 2 accurate results, Satori recorded a larger accuracy in single entity type result than GKG with 61 accurate results while GKG having 26 accurate result.

For the Geo-queries, Satori recorded an overall of 19.73% accurate (labeled 1) result which means that 80.27% of the queries were inaccurate (labeled 0), 61.34% of the queries have no result from Satori and 18.93% of the Geo-queries have results that either mismatch the query type or incorrect results.

From the results above, Satori has a slightly better result than GKG with an overall accurate result of 19.73% while GKG has 18.13%. Both search engines don't return result about statistical queries and attribute of multiple entities. GKG has much better result for multiple entities typed queries while Satori shows a better result for single entity typed queries.

5.6.2 Geo-Queries Result Based On Their Complexities

5.6.2.1 GKG Result

The results for testing the Geo-queries in GKG by considering the complexity of the queries are shown in Figure 5.7 (for single entity type), Figure 5.8 (for multiple entity type), Figure 5.9 (for attribute of an entity type), Figure 5.10 (for attribute of multiple entities type) and Figure 5.11 (for statistical query type).

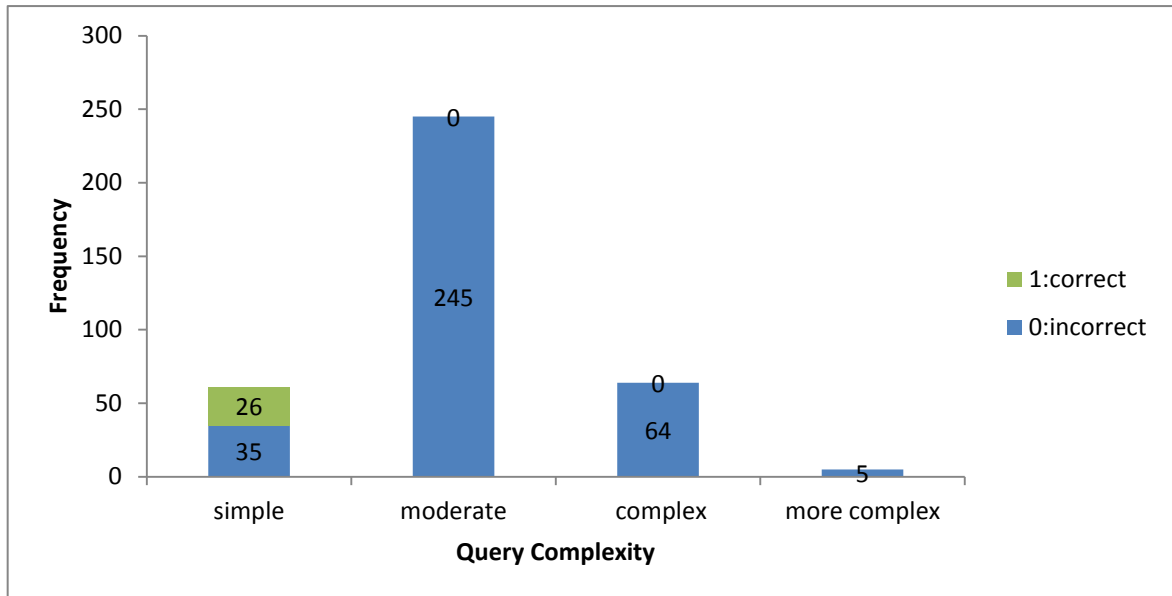


Figure 5.7: GKG result for single entity type from the Geo-queries.

For the single entity type result (Figure 5.7), there is a total of 375 queries, 61 of which are simple, 245 moderate, 64 complex and 5 more complex queries. From the figure, it shows that only in the simple queries GKG was able to respond to the queries correctly. All of the queries belonging to the moderate, complex or more complex complexity usually have no results in GKG. For the single entity type, GKG was able to correctly get 42.62% of the simple queries. GKG could correctly output the answer to the query “What is the capital of Texas?” but couldn’t give the correct answer for the query “Can you tell me the capital of Texas?”

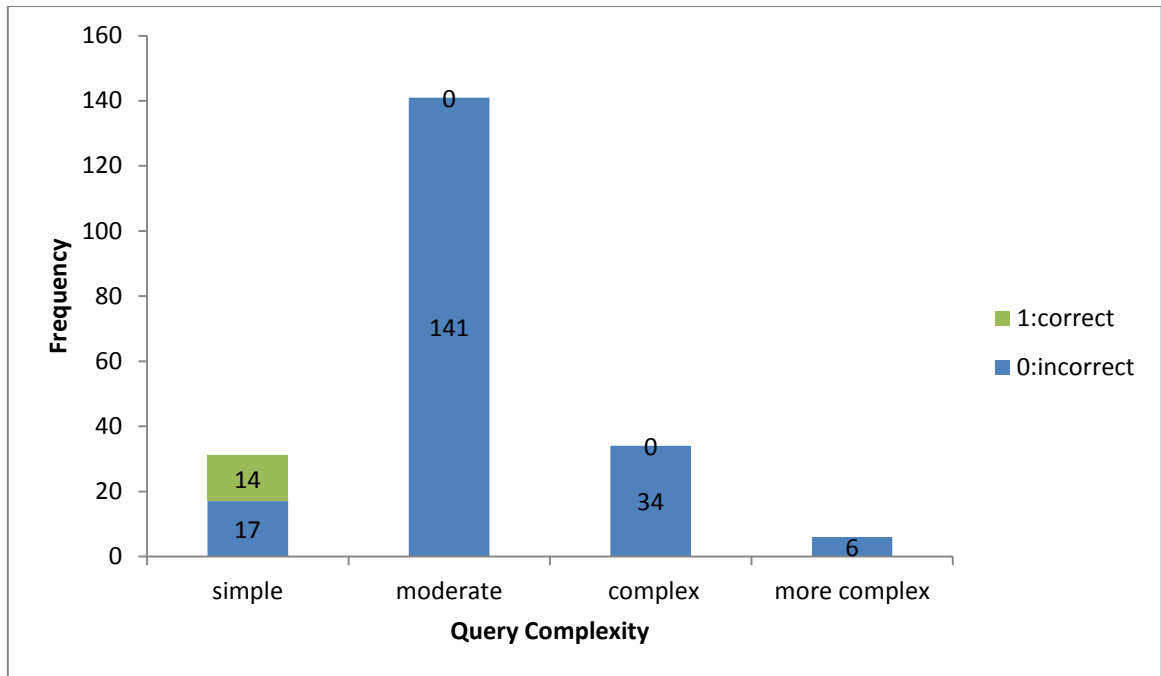


Figure 5.8: GKG result for multiple entities type from the Geo-queries.

The result of multiple entity type (Figure 5.8) is similar to the single entity type. Also only the simple queries have correct results in GKG. About 45% of the simple queries were correctly outputted by GKG. There are 31 simple queries, 141 moderate, 34 complex and 6 more complex queries. GKG correctly outputted the result for the queries “Name the rivers in Arkansas?” and “Rivers in New York?” but couldn’t get the correct result for the queries “Give me all the states of USA?” and “Name all the lakes of us?” which are all simple queries.

GKG has classes implemented for rivers of a state, lakes of a state and mountains of a state. It could output all the rivers, lake and mountains of a state for example “what rivers are in New Mexico?” But it could not output results for queries that should satisfy some conditions or perform some operation. Example queries are “What are major rivers in Texas?” and “What rivers are in states that border Texas?”. Conversely, GKG are yet to have classes implemented for cities of a state, states bordering other states, neighboring states and states adjoining other states. GKG does not return results for queries like “Give me the cities in Virginia?” and “what states border Florida?”.

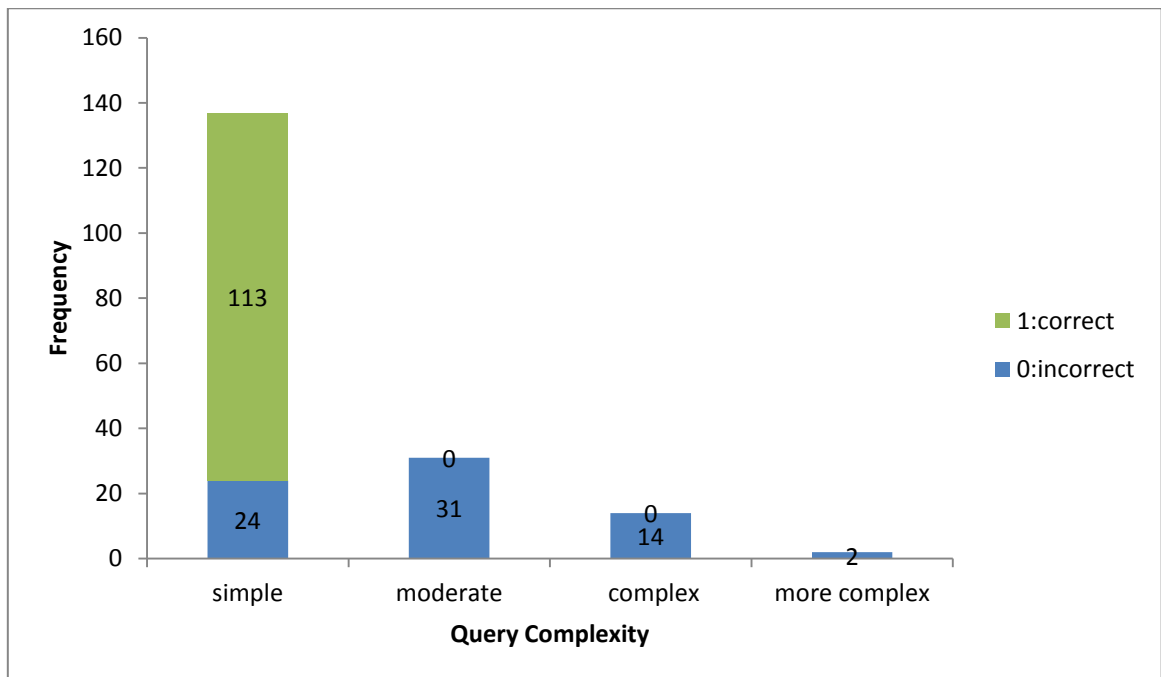


Figure 5.9: GKG result for attribute of an entity type from the Geo-queries.

In the attribute of an entity type result (Figure 5.9), GKG has a high percentage of correct results in the simple queries having gotten 82.48% of the simple queries correctly. The result from the attribute of an entity type is also similar to the single entity and multiple entity type.

The way queries are expressed can affect the results in search engines. Queries that looks similar to people may not be viewed similar in search engines. GKG does not understand the query “How many citizens in Alabama?” or “How many residents live in Texas?” although it returns the answer for the “What is the population of Atlanta?” and “How many people live in Montana?”. The search engine can predict the area of a state and the population of a state or city but could not calculate the population density of states and cities.

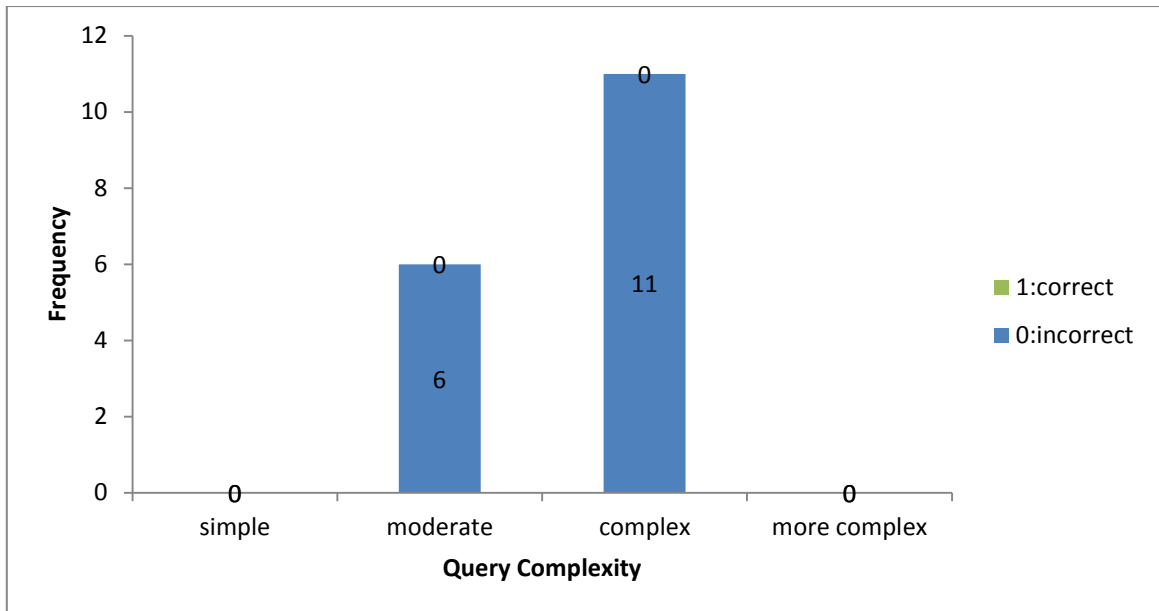


Figure 5.10: GKG result for attribute of multiple entities type from the Geo-queries.

In the attribute of multiple entity type (Figure 5.10), the queries consist of only moderate and complex queries. In both queries, GKG has no convincing output with all of the queries having no result from GKG.

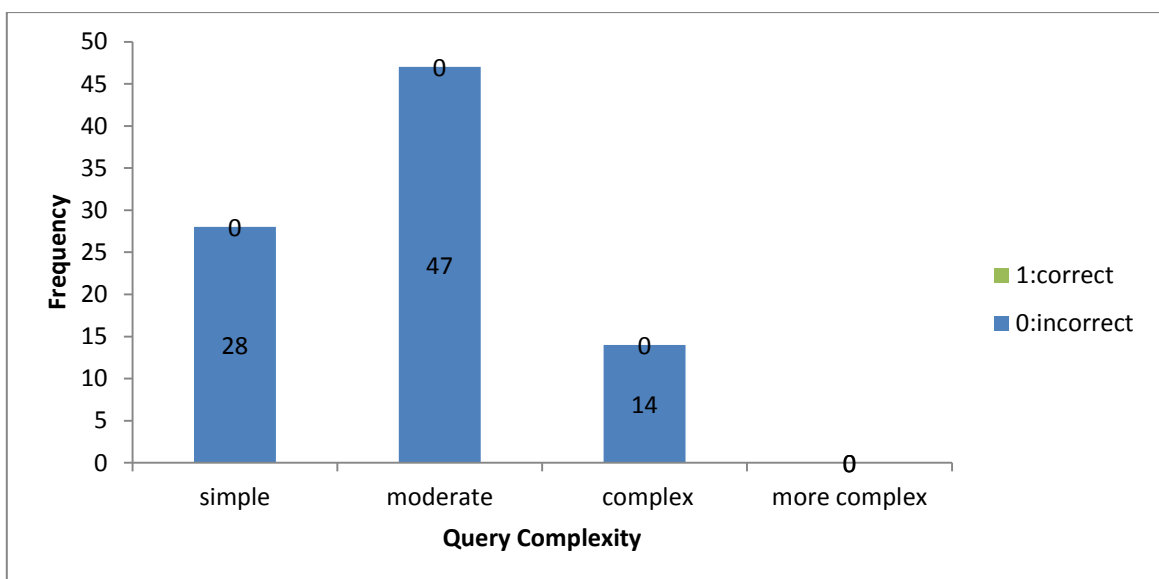


Figure 5.11: GKG result for statistical queries from the Geo-queries.

The statistical queries composed of simple, moderate and complex queries. Neither of these queries could GKG outputted a correct result even with some simple queries like “How many cities are there in USA?” or “How many rivers are there in US?”

The result for the combine query types is shown in Figure 5.12. The result shows that GKG could only output correct results for only simple queries for all the query types. Even among the simple queries, GKG could answer only 158 out of 272 of the simple queries which is equivalent to 58.08% of the simple queries. GKG could not give any correct result for all of the 455 queries with moderate complexities, 137 complex queries and 13 more complex queries.

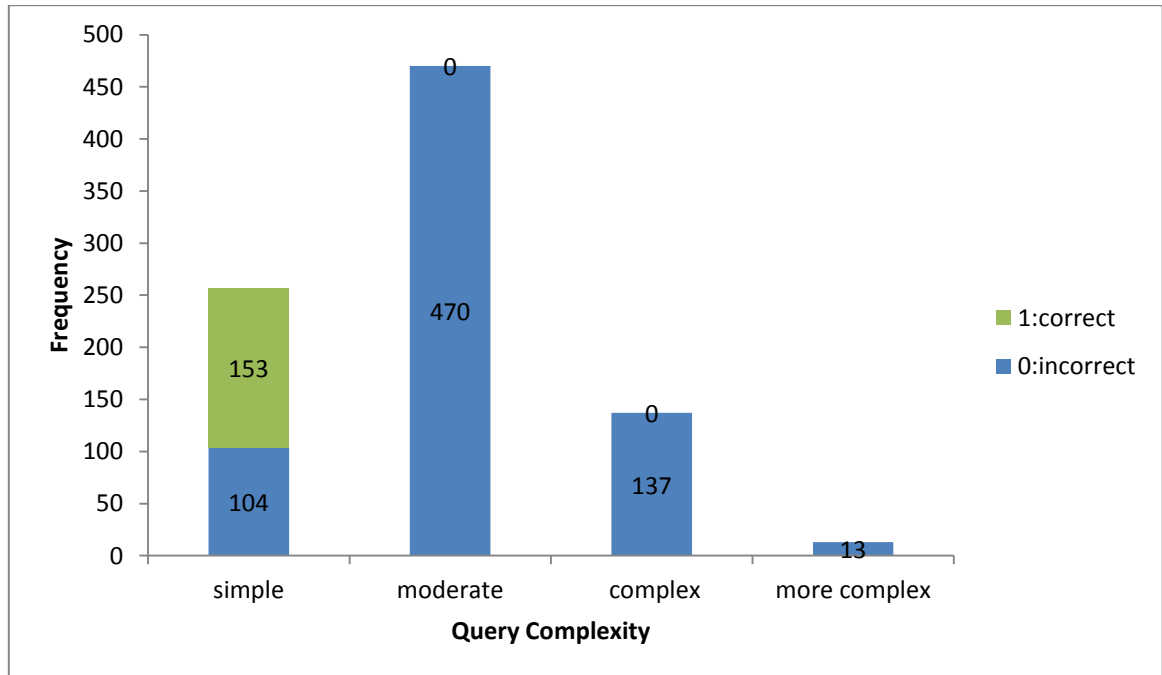


Figure 5.12: GKG result with respect to query complexities.

5.6.2.2 Satori Result

The results for testing the Geo-queries in GKG by considering the complexity of the queries are shown in Figure 5.13 (for single entity type), Figure 5.14 (for multiple entity type), Figure 5.15 (for attribute of an entity type), Figure 5.16 (for attribute of multiple

entities type) and Figure 5.17 (for statistical query type). The results for Satori are almost similar with GKG except for the single and multiple entity types. As we expected based on our previous results from other chapters, Satori performs much better for single entity type queries. Satori could give correct answer for some simple, moderate and complex queries for single entity type queries while GKG could only output correct results for simple queries of single entity type. However, GKG performs much better than Satori for multiple entity type queries. GKG answered 14 out of 31 (45%) of the simple queries of multiple entity types, while Satori could only answer 2 (6%) of these queries correctly. Conversely, neither of the search engines could output the correct answers for queries with complexities other than simple.

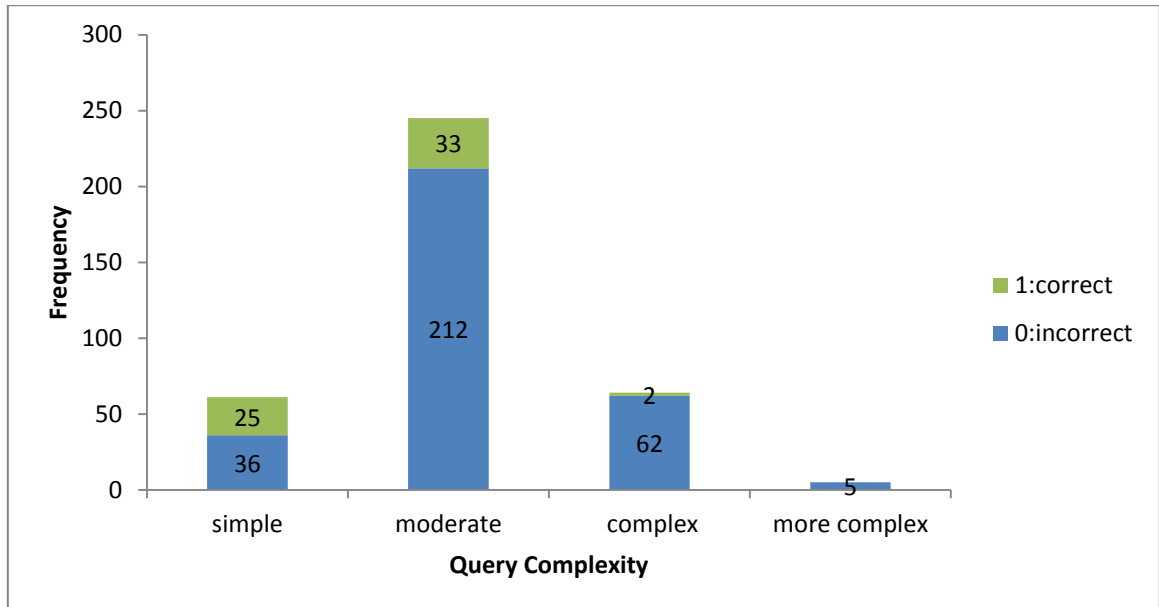


Figure 5.13: Satori result for single entity type from the Geo-queries.

From Figure 5.13, it shows that Satori answered 25 of the simple queries, 33 of the moderate queries, 2 of the complex queries and none of the more complex queries in the single entity type queries which is equivalent to 40.98%, 13.47%, 3.12% and 0% respectively. As we expected, the result shows that as the query complexities increases, the search engines are likely to give no results for the queries.

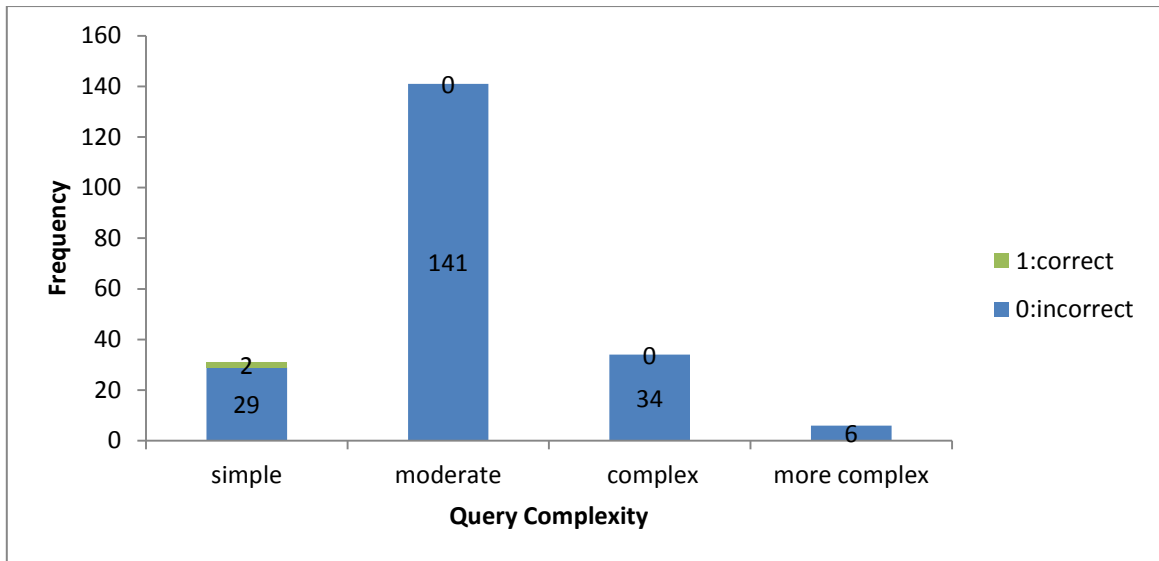


Figure 5.14: Satori result for multiple entities type from the Geo-queries.

The result of multiple entity type in Satori is shown Figure 14. From the Figure, there are 31 simple queries, 141 moderate, 34 complex and 6 more complex queries. Satori GKG could only answer 2 of the simple queries and none for other query complexities. Satori correctly outputted the result for the queries “Rivers in New York?” and “which rivers are in Alaska?” but couldn’t get the correct result for the queries “what are the rivers in Alaska?” as of February, 2014.

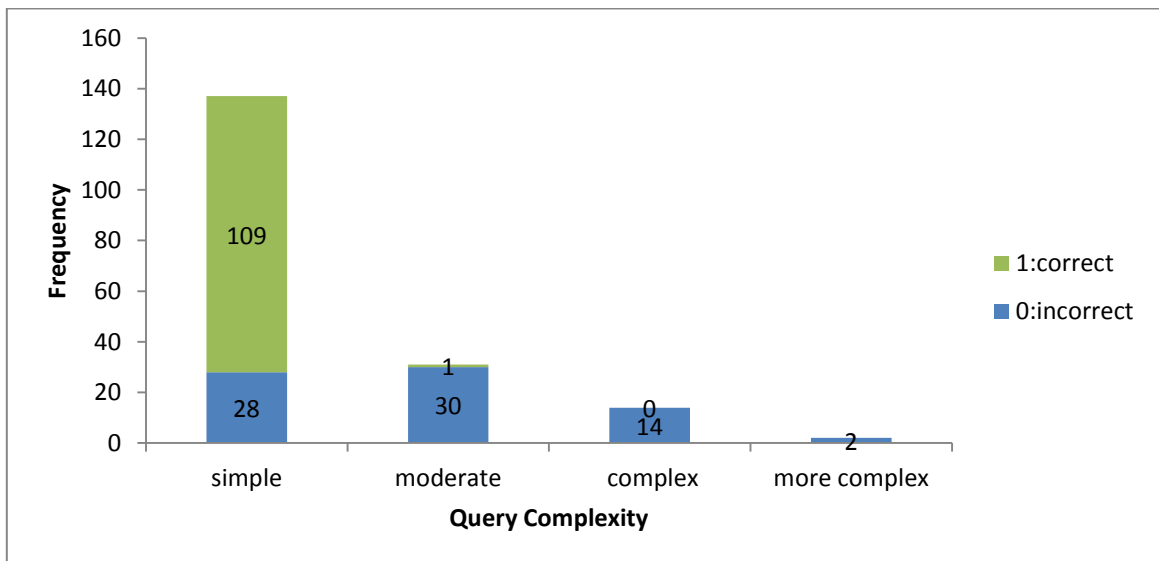


Figure 5.15: Satori result for attribute of an entity type from the Geo-queries.

Satori result for attribute of an entity type (Figure 5.15) is similar to that of GKG. Satori has about 80% of the simple queries correctly but only got 1 of the moderate queries and none of the complex and more complex queries.

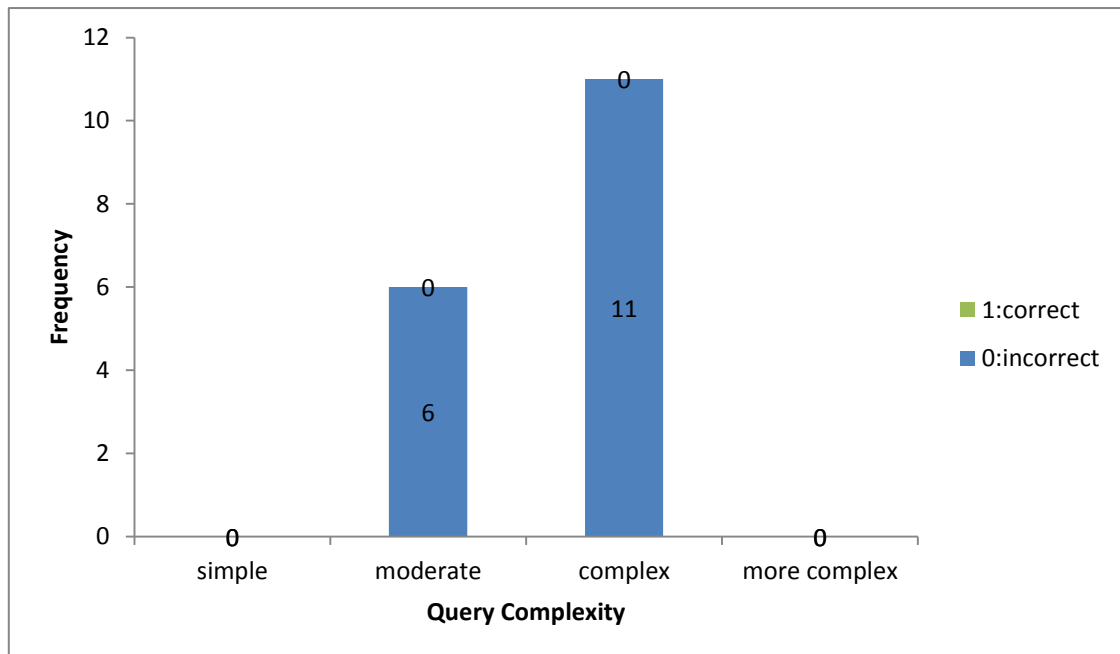


Figure 5.16: Satori result for attribute of multiple entities type from the Geo-queries.

In the attribute of multiple entity type (Figure 5.16), the queries consist of only moderate and complex queries. In both queries, Satori has no encouraging result with all of the queries having no result from Satori.

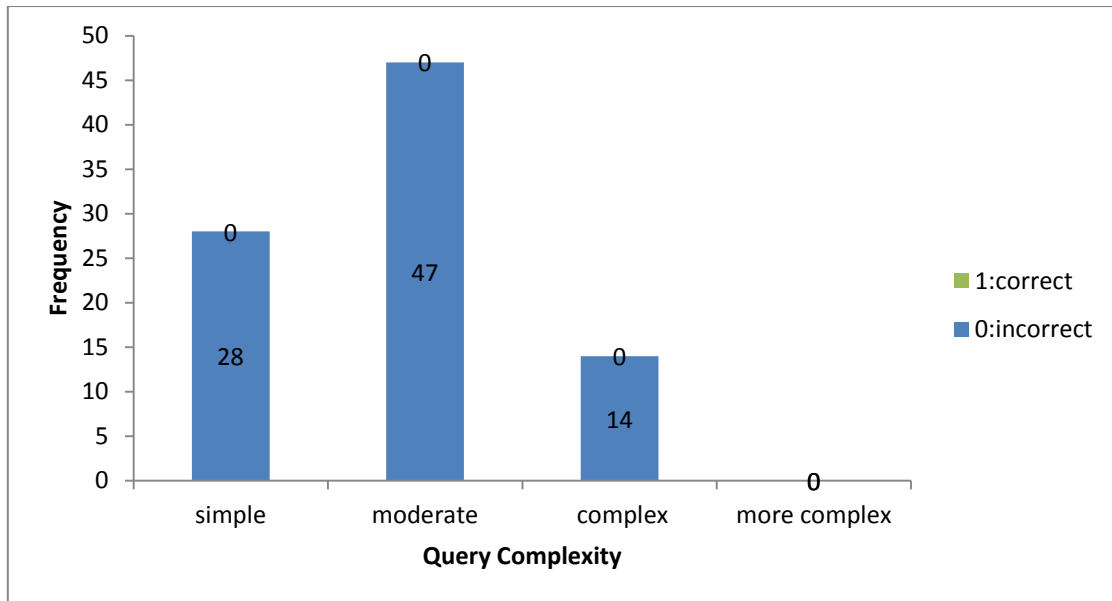


Figure 5.17: Satori result for statistical queries from the Geo-queries.

The result for the statistical queries in Satori (Figure 5.17) is similar to that of GKG. Neither of the queries could Satori or GKG output a single correct result even with some simple queries like “How many cities are there in USA?” or “How many rivers are there in US?”

Satori result for the combine query types is shown in Figure 5.18. From the figure, Satori answer 52.92% of the simple queries for all query types, 7.23% for moderate queries, 1.46% for complex queries and 0% for more complex queries.

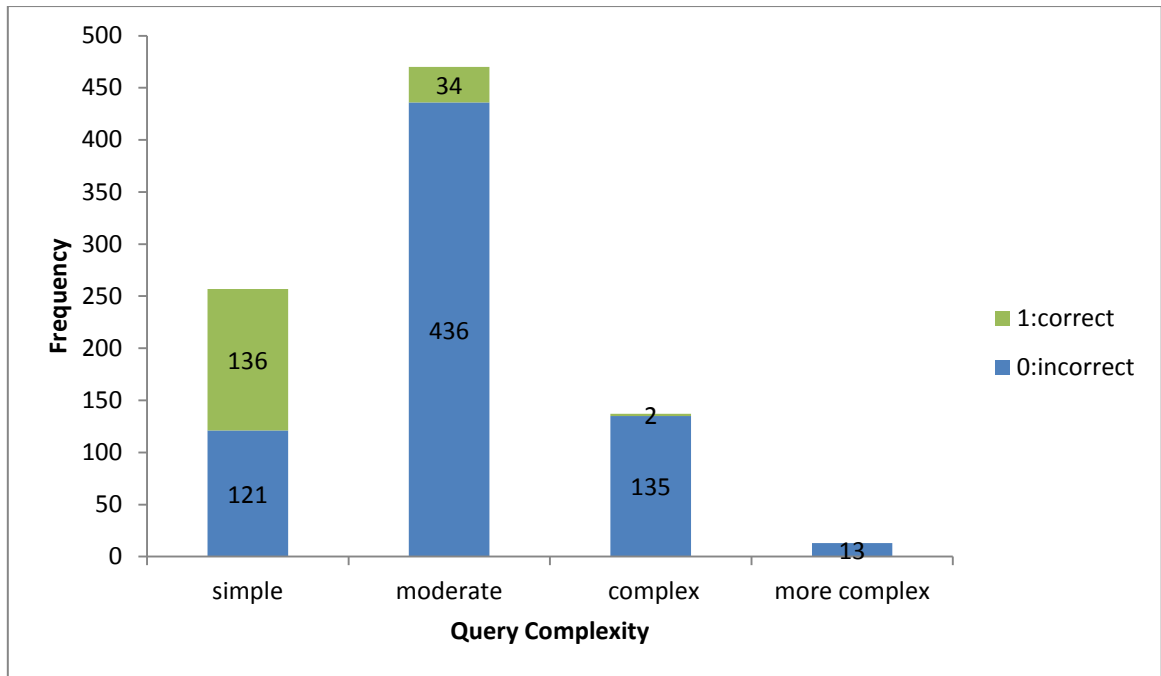


Figure 5.18: Satori result based on query complexities for combine query types.

5.6.3 Yahoo Queries Result

For the Yahoo queries, we expected a much better response from both search engines than in the Geo-queries; since the queries were made of fewer words which are mostly the name of the entities being searched. Moreover, we expect the search engines to suffer less from the habitability problem since the fewer words in the query will give the search engine a better understanding of the user intent (more words mean more confusion to the search engines).

The result of the evaluation from the Yahoo! Queries for GKG and Satori are shown in Figure 10 and Figure11 respectively.

5.6.3.1 GKG Result

The result in Figure 5.19 shows that GKG has 53.02% of the queries tested. This means that about 46.98% of the queries have either no result in GKG or incorrect output by GKG. Among the 46.98% of the queries marked incorrect in GKG, only one query GKG

was able to give an answer, the rest of the queries have no result in GKG. GKG suggested “West Surrey Racing” for the query “ebay motors”. GKG could not give results for some queries like: “google search engine”, “wikipedia”, “avg”, “espn sports”, “msn messenger”, “internet explorer”, “java”, “sky sports” etc.

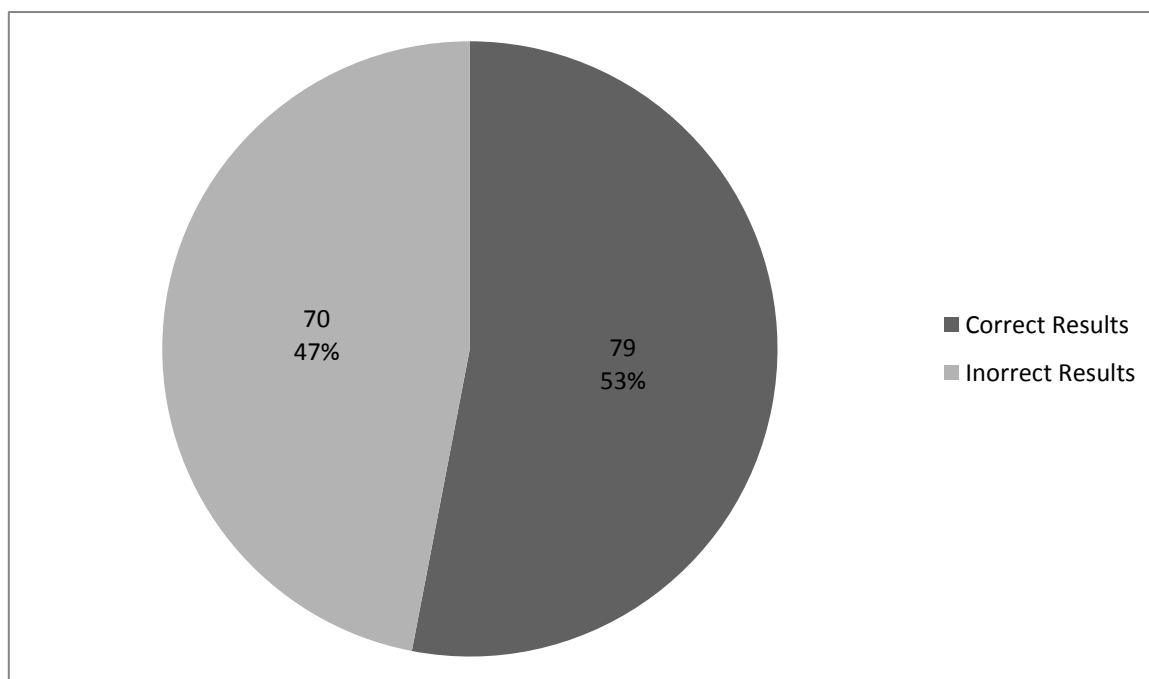


Figure 5.19: GKG result for the Yahoo queries

5.6.3.2 Satori Result

The result from Satori is much better Figure 5.20. This is in consistent to the result from the Geo-queries. Satori has more number of correct results in the single entity type queries. From the Figure, Satori has about 80.5% of the queries correctly. Among the queries marked incorrect, only three were wrongly suggested by Satori and the rest has no result in Satori. Satori outputted “Yahoo!” for the query “yahoo games”, “Yellowpages.com” for the query “yahoo yellow pages”, “ebay” for the query “ebay motors”. Satori could not give any output for queries like: “fling”, “flowers”, “ultimate guitar”, “ovguide”, “voyeur”, “isohunt” etc.

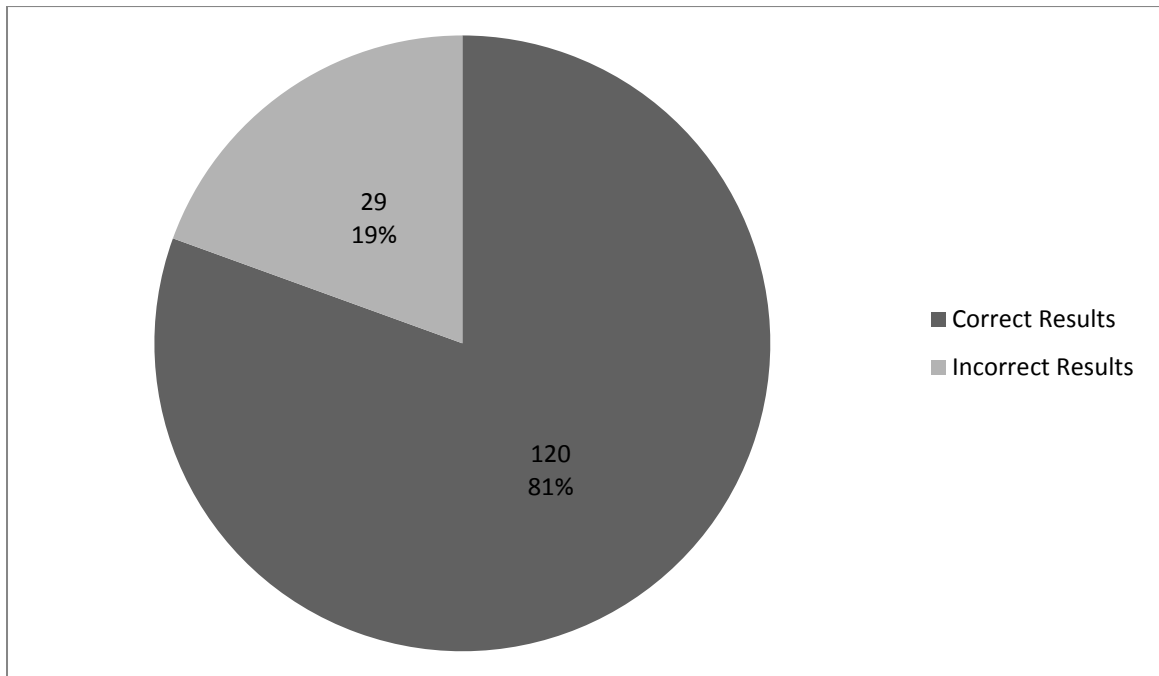


Figure 5.20: Satori result for the Yahoo queries

5.6.4 Yahoo Query Result For List Search

We also examine list search queries (queries about list of entities) in the Yahoo queries. Our investigation show that only few percent of the queries are asking of list of entities. Only 43 out of 1000 (4.3%) of the queries are searching for multiple entities. Most of the list search queries are not specific which means that the search engines needed additional processing or information to understand what exactly the user is seeking. For example, for queries like “car insurance”, “cars”, “movies”, “hotels”, “games” it will be difficult for the search engines to output list of entities for the queries. The queries are about an entire class. In this case the user needs to be specific to fetch entities from the search engines by adding some constraints. In the first query, the user is likely looking for car insurance companies. Query like “top car insurance companies” or “top car insurance companies in USA” would have been better. A refine of the other queries would be “2013 car models”, “movies by Will Smith” or “2013 action movies”, “hotels in New York”, “2013 video games”.

Among the 43 queries we tested, neither of the search engines could output a single list of entities for the queries. We predicted two reasons behind this:

- Poor query: as we have explained above, some queries may not be good enough to give an answer in the search engines, even though we have guaranty that some of the entities users request from the search engines are indexed. Users need to be specific to the list of entities they are seeking for and not the entire entity class like movies, hotels, cars, games, flowers, videos etc.
- Unindexed entity types: some of the entities users are seeking may not have been indexed by the search engines. For example, it is not certain if GKG and Satori have indexed entities about the queries: “area codes”, “baby names”, “cartoon network games”, “hairstyles”, “jobs”, “love poems”, “love quotes”, “wedding dresses”, “zip codes” etc.

In some cases, the search engines may predict some entities in the queries and conclude to return them as the answers to the queries. This is notice especially in Satori. For example, Satori outputted ‘Honda’ for the query “Honda motorcycles”, “Yahoo” for the query “yahoo personals”, ‘monster’ for the query “monster jobs” and ‘miniclip’ for the query “miniclip games”.

5.7 CONCLUSIONS

Semantic search engines that use natural language interfaces still suffer from habitability problem. It shows that semantic search engines are strongly affected by the sentences users used to express their queries. Parsing and understanding user’s queries has been difficult, expensive and time consuming.

Our result on the investigation of queries supported by GKG and Satori shows that the semantic engines could give results about three types of queries (single entity, multiple entities and attribute of entities). The result also shows that both GKG and Satori does not provide results for statistical queries about entities in their dataset such as “how many states are in USA” or “how many rivers are in USA”. The search engines could not also give

results about attribute of multiple entities such as “what are the population densities of each US state” or “What are the highest points of all the states”. Our result in this chapter also validates our findings in chapter three and four. It shows that Satori performs better for searches about single entities while GKG performs better for searches about multiple entities.

A vast majority of the Geo-queries has no result in both search engines. 74.91% of the Geo-queries have no result in GKG while 61.34% of the Geo-queries have no result in Satori. Also 81.87% of the queries have either no result or incorrect result in GKG while 80.27% of the queries have no result or incorrect result in Satori. These large amounts of figures show that both of the search engines do not understand a lot of the Geo-queries which are natural language queries. The search engines suffer from habitability problem (the mismatch between user’s expectation and the capability of natural language understanding of the system). For example as of the time we tested these queries, both GKG and Satori returned ‘Alabama’ for the query “what are the major cities in Alabama”.

Our result in Figure 5.12 and Figure 5.18 also shows that both of the search engines are affected by the complexity of queries. From our result, it shows that GKG could not give correct results for queries other than simple queries. On the other hand, Satori could give result for some moderate query complexity and only 2 of the complex queries in addition to some simple queries. This indicates that GKG and Satori are affected by the complexity of user’s queries. As the complexity of the queries increases from simple to more complex queries, the expectation of users getting result in the search engines decline irrespective of the type of query type. The queries in the Geo-queries may not be the kinds of queries users query search engines with, but the idea is to test how far the engines adopt to query complexity.

However, a test of how the search engines respond to user’s queries about what they frequently ask on the web shows that the search engines performance is encouraging. The result of testing the most frequently used 1000 queries of 2008 from Yahoo shows that the search engines are of significant important in answering user’s queries in the real sense. A

vast majority of the queries users ask are entity centric and of single entity type. Comparing the result in Figure 5.19 and Figure 5.20, it shows that Satori answer more of the queries users frequently ask than GKG. 81% of the queries was answered by Satori while GKG trends with 53%. This is understandable because most of these queries are single entity seeking queries and for all of our test about single entity search satori happen to answer more than GKG. It is possible that Satori may have more entities in their entity database than GKG. However, this does not say anything of the quality of the result the search engines outputted. Conversely, GKG performs better in outputting results about multiple entity type queries. It therefore shows that GKG process more of its entity set than Satori.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this chapter, we summarized our finding for all the evaluations carried out on the semantic search engines and finally our plans for the future research.

6.1 CONCLUSIONS

Our investigation on the kind of entity types in the semantic search engines have shown that most of the entity types in these semantic search engines are of things that are very popular and most interested by people such as: movies, songs, celebrities, theaters, tourist attractions, museums, rivers, lakes, waterfalls, houses, hotels, books, academic institutions, sport teams, companies etc. The kinds of classes they do not index includes: academic degrees, infectious diseases, bones, veins, nerves, anatomical structures, manufactured drug form, Olympic games, Football world cup, Satellite Type, Calendar System, Unit of Data Transmission Rate, File Format, Student radio station etc. From our result, Satori has 66% of the classes we tested while GKG has 60% of the classes we tested.

Our investigation has also shown that GKG and Satori may have fewer classes than some of the entity dataset like Probase, Yago, and Cyc which have thousands to millions of entity types. The reason behind this may be as a result of human intervention in selecting what kind of entity types to be included or not included in their dataset rather than having a particular algorithm that automatically find and index new entity types in the dataset, regardless of what type it is. The human intervention is aimed at providing a high accuracy to satisfy users with the appropriate answers for their query. If entity types are to be

extracted and indexed algorithmically or automatically, as it is done in most document search engines, GKG and Satori will have a large amount of entity types in their dataset, but may tend to be noisy given the vastly different content in the large collection of sites [19]. Therefore, the information stored in this case may not meet to users demand.

Our investigation on the entity type organization among the semantic search engines have shown that both GKG and Satori do not show taxonomical hierarches of entity types to their users, based on what they displayed in their SERP. They may not have taxonomical hierarchy defined in their entity databases in which one class is able to extend the other to inherit the class properties. They may be similar to Freebase data organization in which there is no hierarchy among entity types but related concepts are grouped in the same domain. In both GKG and Satori, hierarchies are shown in very few classes such as animal, drugs and the person class. But these hierarchies are not based on the taxonomical scenario, since there is no sign of inheritance such that a class can inherit all of the features from its superclass. The hierarchies in this case are implemented as features and there is no is-a relationship between entity types but users can view and navigate to higher or lower classification.

We also find out that entity of the same type may have different attributes defined and different attributes may be displayed for entities of the same type as results in the semantic search engines, even though they may have those properties defined. For example, different properties may be returned for two different scientists or professors such as Larry Page and Sergey Brin. The fact is that what users are interested about one entity may not be the case in another entity of the same type.

Our investigation of list search services by the semantic engines shows that the list search service are implemented only for few classes in both engines. The kinds of classes they display list of entities includes: Amusement Rides, Houses, Academic institutions, Musical Groups, Person, Theaters, Tourist attractions, American football team, Animal breeds, Geological formations and Planets. GKG returned list of entities for 10 classes out of 51 classes we tested which is about 20% of the classes we tested. Satori returned list

result for 14% of the classes tested. Our result also shows that GKG has implemented list search services for only 17% of its classes while Satori has implemented only 11% of list search services among their entity types.

An investigation of the queries supported by the semantic engines has shown that both GKG and Satori could answer some queries about single entity, multiple entities and attribute of entities. But they could not give single result for statistical queries and queries seeking for attributes of more than one entity. Satori answered more of the single entity queries while GKG answered more of the multiple entity queries. However, their performances on attribute of entity queries seem to only differ slightly.

Our investigation on the natural language understanding capabilities of the semantic search engines shows that the search engines suffer from habitability problem (the mismatch between user's expectation and the capability of natural language understanding of the system). For example as of the time we tested these queries, both GKG and Satori returned 'Alabama' for the query "what are the major cities in Alabama". A vast majority of the Geo-queries which are natural language queries has no result in both search engines. 74.91% of the Geo-queries have no result in GKG while 61.34% of the Geo-queries have no result in Satori. Also 81.87% of the queries have either no result or incorrect result in GKG while 80.27% of the queries have no result or incorrect result in Satori. These large amounts of figures show that both of the search engines do not understand most of natural language queries.

Our result also indicates that GKG and Satori are affected by the complexity of user's queries. As the complexity of the queries increases from simple to more complex queries, the expectation of users getting result in the search engines decline irrespective of the type of query type. The queries in the Geo-queries may not be the kinds of queries users query search engines with, but the idea is to test how far the engines adopt to query complexity.

However, a test of how the search engines respond to user's queries about what they frequently ask on the web shows that the semantic engines performance are encouraging.

The result of testing the most frequently used 1000 queries of 2008 from Yahoo shows that the search engines are of significant importance in answering user's queries in the real sense. A vast majority of the queries users ask are entity centric and of single entity type. Comparing the result in Figure 5.19 and Figure 5.20, it shows that Satori answers more of the queries users frequently ask. 81% of the queries were answered by Satori while GKG answers 53% of the queries. This is understandable because most of these queries are single entity seeking queries and for our entire test about single entity search Satori happens to answer more the single entity queries. It is possible that Satori may have more entities in their entity database than GKG. However, this does not say anything of the quality of the result the search engines outputted.

It is important to note that the result of our evaluation is likely to change over time as the semantic engines index more and more entities in their database. We expect to see more improvements in the quality and the frequencies in which results are shown in the semantic search engines in the future.

6.2. FUTURE WORK

In the future work, we would like to study the accuracy of the semantic search engines and the quality result returned by the semantic search engines. Since they choose which properties to return to their users, how convincing are the results returned in the user's perspective. Also are the results returned by the semantic search engines enough to answer users' needs? This will be interesting to investigate.

We hope to also investigate class coverage among the semantic search engines also. In the class coverage, we will investigate how much of the entities of a class are recognized or present in either of the semantic search engines we used in this research. For example, how much of people, universities, celebrities, buildings or geographical features are recognized in either Google Knowledge Graph or Satori.

It is possible to repeat the evaluation also to monitor how fast the semantic engines are growing and changing. It's most likely that the result we obtained could change over time as GKG and Satori index more entities and improve their search algorithms.

REFERENCES

- [1] Google, <http://www.google.com> (Accessed on 17 January, 2014).
- [2] Yahoo, <http://www.yahoo.com> (Accessed on 17 January, 2014).
- [3] Bing, <http://www.bing.com> (Accessed on 17 January, 2014).
- [4] Yandex, <http://www.yandex.com> (Accessed on 17 January, 2014).
- [5] Hakia, <http://www.hakia.com> (Accessed on 17 January, 2014).
- [6] Ding, Li, et al. "Swoogle: a search and metadata engine for the semantic web." Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004.
- [7] Doms, Andreas, and Michael Schroeder. "GoPubMed: exploring PubMed with the gene ontology." Nucleic acids research 33.suppl 2 (2005): W783-W786.
- [8] Rajaraman, Anand. "Kosmix: high-performance topic exploration using the deep web." Proceedings of the VLDB Endowment 2.2 (2009): 1524-1529.
- [9] Matuszek, Cynthia, et al. "An Introduction to the Syntax and Content of Cyc." AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering. 2006.
- [10] Wu, Wentao, et al. Towards a probabilistic taxonomy of many concepts. Technical Report MSR-TR-2011-25, Microsoft Research, 2011.
- [11] Bollacker, Kurt, et al. "Freebase: a collaboratively created graph database for structuring human knowledge." Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008.
- [12] Freebase, <http://www.freebase.com> (Accessed on 17 January, 2014).
- [13] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

- [14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *The semantic web*, pages 722{735, 2007.
- [15] Amit Singhal, *Introducing the Knowledge Graph: things, not strings*, Google Blog post, May 16, 2012, <http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>, accessed Sep 16th, 2013.
- [16] Richard Qian, *Understand Your World with Bing*, Bing Blog Website, March, 2013, http://www.bing.com/blogs/site_blogs/b/search/archive/2013/03/21/satorii.aspx, accessed Sep 16th, 2013.
- [17] Amir Efrati, “Google Gives Search a Refresh”, *Wall Street Journal*, March 15, 2012, <http://online.wsj.com/article/SB10001424052702304459804577281842851136290.html>
- [18] David Pogue, “Going Beyond Search, Into Fetch”, *New York Times*, May 23, 2012, <http://www.nytimes.com/2012/05/24/technology/personaltech/google-and-microsoft-feature-do-it-all-search-pages-state-of-the-art.html>
- [19] Dalvi, Nilesh, et al. "A web of concepts." *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2009.
- [20] Prud'hommeaux, E., Seaborne, A.: *SPARQL Query Language for RDF*. W3C Recommendation (January 2008) <http://www.w3.org/TR/rdf-sparql-query/>.
- [21] Neo4j Team, *Cypher Query Language*, <http://docs.neo4j.org/chunked/stable/cypher-query-lang.html>, Accessed on Nov 20th, 2013.
- [22] Damjanovic, D., Agatonovic, M., & Cunningham, H. (2010). Natural language interfaces to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction. In *The Semantic Web: Research and Applications* (pp. 106-120). Springer Berlin Heidelberg.
- [23] Kaufmann, E., & Bernstein, A. (2010). Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4), 377-393.
- [24] Sergey Brin and Larry Page. *The anatomy of a large-scale hypertextual web search engine*. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.
- [25] Jon Kleinberg. *Authoritative sources in a hyperlinked environment*. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [26] Benczur, A. A., Csalogany, K., Sarlos, T., & Uher, M. (2005, May). *SpamRank—Fully Automatic Link Spam Detection Work in progress*. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*.

- [27] Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1), 1.
- [28] Kaufmann, E., Bernstein, A., Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases, *Journal of Web Semantics*, Vol. 8 (4), 2010.
- [29]] E. Kaufmann, A. Bernstein, R. Zumstein, Querix: A Natural Language Interface to Query Ontologies Based on Clarification Dialogs, in: 5th International Semantic Web Conference (ISWC 2006), Athens, GA, 2006.
- [30] A. Bernstein, E. Kaufmann, C. Kaiser, Querying the Semantic Web with Ginseng: A Guided Input Natural Language Search Engine, in: 15th Workshop on Information Technologies and Systems (WITS 2005), Las Vegas, NV, 2005.
- [31] Research Cyc, <http://www.cyc.com/platform/researchcyc>, accessed on Dec 6th
- [32] Saracevic, Tefko. "Evaluation of evaluation in information retrieval." *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995.
- [33] Vaughan, Liwen. "New measurements for search engine evaluation proposed and tested." *Information Processing & Management* 40.4 (2004): 677-691.
- [34] Tumer, Duygu, Mohammad Ahmed Shah, and Yiltan Bitirim. "An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia." *Internet Monitoring and Protection, 2009. ICIMP'09. Fourth International Conference on*. IEEE, 2009.
- [35] Steiner, Thomas, and Stefan Mirea. "SEKI@ home, or Crowdsourcing an Open Knowledge Graph." *Proceedings of the First International Workshop on Knowledge Extraction and Consolidation from Social Media (KECSM2012)*, Boston, USA. 2012.
- [36] Steiner, Thomas, et al. "Adding Realtime Coverage to the Google Knowledge Graph." *International Semantic Web Conference (Posters & Demos)*. 2012.
- [37] Henzinger, Monika R. "Algorithmic challenges in web search engines." *Internet Mathematics* 1.1 (2004): 115-123.
- [38] Lewandowski, Dirk. "Challenges for Search Engine Retrieval Effectiveness Evaluations: Universal Search, User Intents, and Results Presentation." *Quality Issues in the Management of Web Information*. Springer Berlin Heidelberg, 2013. 179-196.
- [39] Buckley, Chris, and Ellen M. Voorhees. "Retrieval evaluation with incomplete information." *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004.

- [40] Wrigley, Stuart N., et al. "Evaluating semantic search tools using the SEALS Platform." International Workshop on Evaluation of Semantic Technologies (IWEST 2010), International Semantic Web Conference (ISWC2010), International Semantic Web Conference (ISWC2010), China. 2010.
- [41] Mäkelä, Eetu. "Survey of semantic search research." Proceedings of the seminar on knowledge management on the semantic web. Department of Computer Science, University of Helsinki, Helsinki, 2005.
- [42] Tumer, Duygu, Mohammad Ahmed Shah, and Yiltan Bitirim. "An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia." Internet Monitoring and Protection, 2009. ICIMP'09. Fourth International Conference on. IEEE, 2009.
- [43] Blanco, Roi, et al. "Entity search evaluation over structured web data." Proceedings of the 1st international workshop on entity-oriented search workshop (SIGIR 2011), ACM, New York. 2011.
- [44] Uyar, Ahmet. "Investigation of the accuracy of search engine hit counts." Journal of Information Science 35.4 (2009): 469-480.
- [45] Elbedweihi, Khadija, et al. "Evaluating semantic search systems to identify future directions of research." Proc. 2nd International Workshop on Evaluation of Semantic Technologies (IWEST 2012). 2012.
- [46] Menczer, Filippo, et al. "Evaluating topic-driven web crawlers." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.
- [47] Menczer, Filippo, Gautam Pant, and Padmini Srinivasan. "Topical web crawlers: Evaluating adaptive algorithms." ACM Transactions on Internet Technology (TOIT) 4.4 (2004): 378-419.
- [48] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30.1 (1998): 107-117.
- [49] Page, Lawrence, et al. "The PageRank citation ranking: Bringing order to the web." (1999).
- [50] Rasolofo, Yves, and Jacques Savoy. "Term proximity scoring for keyword-based retrieval systems." Advances in Information Retrieval. Springer Berlin Heidelberg, 2003. 207-218.
- [51] Schenkel, Ralf, et al. "Efficient text proximity search." String Processing and Information Retrieval. Springer Berlin Heidelberg, 2007.
- [52] <https://schema.org/> (accessed March 20th, 2014.)

- [53] Wang, Yue, et al. Toward topic search on the web. Technical report, Microsoft Research, 2010.
- [54] Zelle, John M., and Raymond J. Mooney. "Learning to parse database queries using inductive logic programming." Proceedings of the National Conference on Artificial Intelligence. 1996.
- [55] Pound, Jeffrey, Peter Mika, and Hugo Zaragoza. "Ad-hoc object retrieval in the web of data." Proceedings of the 19th international conference on World wide web. ACM, 2010.
- [56] Uren, Victoria; Lei, Yuanguai; Lopez, Vanessa; Liu, Haiming; Motta, Enrico and Giordanino, Marina (2007). The usability of semantic search tools: a review. The Knowledge Engineering Review, 22(4), pp. 361–377.
- [57] Tang, Lappoon R., and Raymond J. Mooney. "Using multiple clause constructors in inductive logic programming for semantic parsing." Machine Learning: ECML 2001. Springer Berlin Heidelberg, 2001. 466-477.
- [58] Elbedweihy, Khadija, et al. "Evaluating semantic search systems to identify future directions of research." Proc. 2nd International Workshop on Evaluation of Semantic Technologies (IWEST 2012). 2012.
- [59] Kaufmann, Esther, and Abraham Bernstein. "Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases." Web Semantics: Science, Services and Agents on the World Wide Web 8.4 (2010): 377-393.
- [60] Damljanovic, Danica, Milan Agatonovic, and Hamish Cunningham. "Natural language interfaces to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction." The semantic web: Research and applications. Springer Berlin Heidelberg, 2010. 106-120.
- [61] Kaufmann, E.: Talking to the Semantic Web Natural Language Query Interfaces for Casual End-Users. PhD thesis, University of Zurich (2007).
- [62] John, Tony (March 15, 2012). "What is Semantic Search?".Techulator. Retrieved April 13, 2014.
- [63] Bruno, N., Koudas, N., & Srivastava, D. (2002, June). Holistic twig joins: optimal XML pattern matching. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (pp. 310-321). ACM.
- [64] https://developers.google.com/freebase/guide/basic_concepts (Accessed on 17 January, 2014).

- [65] Sekine, Satoshi, Kiyoshi Sudo, and Chikashi Nobata. "Extended Named Entity Hierarchy." In LREC. 2002.
- [66] Balog, Krisztian, and Robert Neumayer. "Hierarchical target type identification for entity-oriented queries." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 2391-2394. ACM, 2012.
- [67] Sean Gallagher, How Google and Microsoft taught search to "understand" the Web: Inside the architecture of Google's Knowledge Graph and Microsoft's Satori, June 7th, 2012. <http://arstechnica.com/information-technology/2012/06/inside-the-architecture-of-googles-knowledge-graph-and-icrosofts-satori/2/> (accessed Feb 6th, 2014.)
- [68] Thomas J Thompson, The answer is out there (cue X Files music). <http://www.bloomweb.com/the-answer-is-out-there-cue-x-files-music/> (accessed Feb 6th, 2014.)
- [69] Mika, Peter, Edgar Meij, and Hugo Zaragoza. "Investigating the semantic gap through query log analysis." The Semantic Web-ISWC 2009. Springer Berlin Heidelberg, 2009. 441-455.
- [70] Google Tests an Expanded Knowledge Graph Box, July 24, 2012. <http://googlesystem.blogspot.com.tr/2012/07/google-tests-expanded-knowledge-graph.html> accessed March 16th, 2014.
- [71] Mike Blumenthal, Google Local Carousel Display Showing More Frequently, In More Categories, JUNE 14, 2013, <http://blumenthals.com/blog/2013/06/14/google-local-carousel-display-showing-more-frequently-in-more-categories/> , accessed March 20th, 2014.
- [72] Barry Schwartz, Updated Google Knowledge Graph Carousel, Jun 17, 2013 8:08 am, <http://www.seroundtable.com/google-knowledge-graph-carousel-update-16937.html> , accessed March 20th, 2014.
- [73] Amy Gesenhues, Google Knowledge Graph Carousel Sightings Becoming More Frequent Within A Wider Variety Of Searches, Jun 17, 2013 at 10:06am, <http://searchengineland.com/google-knowledge-graph-carousel-sightings-becoming-more-frequent-within-a-wider-variety-of-searches-163562> accessed March 20th, 2014.
- [74] Buckley, Chris, and Ellen M. Voorhees. "Retrieval evaluation with incomplete information." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.
- [75] Blanco, Roi, et al. "Entity search evaluation over structured web data." Proceedings of the 1st international workshop on entity-oriented search workshop (SIGIR 2011), ACM, New York. 2011.

- [76] Barry Schwartz, Google's Knowledge Graph Gains "Carousel," Goes Worldwide In English, Aug 8, 2012 at 1:02pm, <http://searchengineland.com/googles-knowledge-graph-now-worldwide-129948>, accessed March 20th, 2014.
- [77] Knuth, D. E., Morris, Jr, J. H., & Pratt, V. R. (1977). Fast pattern matching in strings. *SIAM journal on computing*, 6(2), 323-350.
- [78] Agarwal, G., Kabra, G., & Chang, K. C. C. (2010, April). Towards rich query interpretation: walking back and forth for mining query templates. In *Proceedings of the 19th international conference on World wide web* (pp. 1-10). ACM.
- [79] Yahoo! Webscope (2009) Yahoo! Webscope dataset ydata-search-queries-multiple-langs-v1_0 [http://research.yahoo.com/Academic_Relations]
- [80] Kotis, K., Papasalouros, A., & Maragkoudakis, M. (2009). Mining Web queries to boost semantic content creation. In B. White (Ed.), *Proceedings of IADIS Conference on WWW/Internet* (pp. 158-162).
- [81] C. W. Thompson, P. Pazandak, H. R. Tennant, Talk to Your Semantic Web, *IEEE Internet Computing* 9 (6) (2005) 75–78.